

Volume 14 Issue 2

February 2023



ISSN 2156-5570(Online)

ISSN 2158-107X(Print)



# Editorial Preface

## *From the Desk of Managing Editor...*

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

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

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

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

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

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

**Thank you for Sharing Wisdom!**

**Kohei Arai**  
**Editor-in-Chief**  
**IJACSA**  
**Volume 14 Issue 2 February 2023**  
**ISSN 2156-5570 (Online)**  
**ISSN 2158-107X (Print)**

# Editorial Board

## Editor-in-Chief

### **Dr. Kohei Arai - Saga University**

*Domains of Research: Technology Trends, Computer Vision, Decision Making, Information Retrieval, Networking, Simulation*

---

## Associate Editors

### **Alaa Sheta**

#### **Southern Connecticut State University**

*Domain of Research: Artificial Neural Networks, Computer Vision, Image Processing, Neural Networks, Neuro-Fuzzy Systems*

### **Domenico Ciuonzo**

#### **University of Naples, Federico II, Italy**

*Domain of Research: Artificial Intelligence, Communication, Security, Big Data, Cloud Computing, Computer Networks, Internet of Things*

### **Doroła Kaminska**

#### **Lodz University of Technology**

*Domain of Research: Artificial Intelligence, Virtual Reality*

### **Elena Scutelnicu**

#### **"Dunarea de Jos" University of Galati**

*Domain of Research: e-Learning, e-Learning Tools, Simulation*

### **In Soo Lee**

#### **Kyungpook National University**

*Domain of Research: Intelligent Systems, Artificial Neural Networks, Computational Intelligence, Neural Networks, Perception and Learning*

### **Krassen Stefanov**

#### **Professor at Sofia University St. Kliment Ohridski**

*Domain of Research: e-Learning, Agents and Multi-agent Systems, Artificial Intelligence, e-Learning Tools, Educational Systems Design*

### **Renato De Leone**

#### **Università di Camerino**

*Domain of Research: Mathematical Programming, Large-Scale Parallel Optimization, Transportation problems, Classification problems, Linear and Integer Programming*

### **Xiao-Zhi Gao**

#### **University of Eastern Finland**

*Domain of Research: Artificial Intelligence, Genetic Algorithms*

# CONTENTS

Paper 1: Organizational Digital Transformations and the Importance of Assessing Theoretical Frameworks such as TAM, TTF, and UTAUT: A Review

*Authors: Bibhu Dash, Pawankumar Sharma, Swati Swayamsiddha*

**PAGE 1 – 6**

Paper 2: Adversarial Sampling for Fairness Testing in Deep Neural Network

*Authors: Tosin Ige, William Marfo, Justin Tonkinson, Sikiru Adewale, Bolanle Hafiz Matti*

**PAGE 7 – 13**

Paper 3: Text2Simulate: A Scientific Knowledge Visualization Technique for Generating Visual Simulations from Textual Knowledge

*Authors: Ifeoluwatayo A. Ige, Bolanle F. Oladejo*

**PAGE 14 – 22**

Paper 4: A Review on Artificial Intelligence in the Context of Industry 4.0

*Authors: Shadi Banitaan, Ghaith Al-refai, Sattam Almatarneh, Hebah Alquran*

**PAGE 23 – 30**

Paper 5: A Machine Learning Enabled Hall-Effect IoT-System for Monitoring Building Vibrations

*Authors: Emanuele Lattanzi, Paolo Capellacci, Valerio Freschi*

**PAGE 31 – 40**

Paper 6: Sequence Recommendation based on Deep Learning

*Authors: Gulsim Rysbayeva, Jingwei Zhang*

**PAGE 41 – 54**

Paper 7: Routing Overhead Aware Optimal Cluster based Routing Algorithm for IoT Network using Heuristic Technique

*Authors: Srinivasulu M, Shiva Murthy G*

**PAGE 55 – 64**

Paper 8: Supply Chain Network Model using Multi-Agent Reinforcement Learning for COVID-19

*Authors: Tomohito Okada, Hiroshi Sato, Masao Kubo*

**PAGE 65 – 69**

Paper 9: An Enhanced MCDM Model for Cloud Service Provider Selection

*Authors: Ayman S. Abdelaziz, Hany Harb, Alaa Zaghloul, Ahmed Salem*

**PAGE 70 – 77**

Paper 10: Dynamic Software Architecture Design for Virtual Rehabilitation System for Manual Motor Dexterity

*Authors: Edwin Enrique Saavedra Parisaca, Solansh Jaqueline Montoya Muñoz, Elizabeth Vidal Duarte, Eveling Gloria Castro Gutierrez, Angel Yvan Choquehuanca Peraltilla, Sergio Albiol Pérez*

**PAGE 78 – 85**

Paper 11: Insights into Search Engine Optimization using Natural Language Processing and Machine Learning

*Authors: Vinutha M S, M C Padma*

**PAGE 86 – 96**

Paper 12: Adaptive Rectified Linear Unit (Arelu) for Classification Problems to Solve Dying Problem in Deep Learning

*Authors: Ibrahim A. Atoum*

PAGE 97 – 102

Paper 13: Espousing AI to Enhance Cost-Efficient and Immersive Experience for Human Computer Interaction

*Authors: Deepak Chaturvedi, Ashima Arya, Mohammad Zubair Khan, Eman Aljohani, Liyakathunisa, Vaishali Arya, Namrata Sukhija, Prakash Srivastava*

PAGE 103 – 111

Paper 14: Implementation of Big Data Privacy Preservation Technique for Electronic Health Records in Multivendor Environment

*Authors: Ganesh Dagadu Puri, D. Haritha*

PAGE 112 – 118

Paper 15: Churn Customer Estimation Method based on LightGBM for Improving Sales

*Authors: Kohei Arai, Ikuya Fujikawa, Yusuke Nakagawa, Ryuya Momozaki, Sayuri Ogawa*

PAGE 119 – 125

Paper 16: Privacy Preservation Modelling for Securing Image Data using Novel Ethereum-based Ecosystem

*Authors: Chhaya S Dule, Roopashree H. R*

PAGE 126 – 135

Paper 17: Music Note Feature Recognition Method based on Hilbert Space Method Fused with Partial Differential Equations

*Authors: Liqin Liu*

PAGE 136 – 143

Paper 18: Hyperparameter Optimization of Support Vector Regression Algorithm using Metaheuristic Algorithm for Student Performance Prediction

*Authors: M. Riki Apriyadi, Ermatita, Dian Palupi Rini*

PAGE 144 – 150

Paper 19: Experimental Analysis and Monitoring of Photovoltaic Panel Parameters

*Authors: Zaidan Didi, Ikram El Azami*

PAGE 151 – 157

Paper 20: Hybrid Feature Selection Algorithm and Ensemble Stacking for Heart Disease Prediction

*Authors: Nureen Afiqah Mohd Zaini, Mohd Khalid Awang*

PAGE 158 – 165

Paper 21: Automatic Extraction of Indonesian Stopwords

*Authors: Harry Tursulistiyono Yani Achsan, Heru Suhartanto, Wahyu Catur Wibowo, Deshinta A. Dewi, Khairul Ismed*

PAGE 166 – 171

Paper 22: Software Effort Estimation through Ensembling of Base Models in Machine Learning using a Voting Estimator

*Authors: Beesetti Kiran Kumar, Saurabh Bilgaiyan, Bhabani Shankar Prasad Mishra*

PAGE 172 – 181

Paper 23: An Effective Heart Disease Prediction Framework based on Ensemble Techniques in Machine Learning

*Authors: Deepali Yewale, S. P. Vijayaragavan, V. K. Bairagi*

PAGE 182 – 190

**Paper 24: Explaining the Outputs of Convolutional Neural Network - Recurrent Neural Network (CNN-RNN) based Apparent Personality Detection Models using the Class Activation Maps**

*Authors: WMKS Ilmini, TGI Fernando*

**PAGE 191 – 197**

**Paper 25: Landmark Recognition Model for Smart Tourism using Lightweight Deep Learning and Linear Discriminant Analysis**

*Authors: Mohd Norhisham Razali, Enurt Owens Nixon Tony, Ag Asri Ag Ibrahim, Rozita Hanapi, Zamhar Iswandono*

**PAGE 198 – 213**

**Paper 26: Performance Comparison of the Kernels of Support Vector Machine Algorithm for Diabetes Mellitus Classification**

*Authors: Dimas Aryo Anggoro, Dian Permatasari*

**PAGE 214 – 219**

**Paper 27: Deep Study of CRF Models for Speech understanding in Limited Task**

*Authors: Marwa Graja*

**PAGE 220 – 226**

**Paper 28: Paw Search – A Searching Approach for Unsorted Data Combining with Binary Search and Merge Sort Algorithm**

*Authors: Md. Harun Or Rashid, Ahmed Imtiaz*

**PAGE 227 – 236**

**Paper 29: A Survey of Forensic Analysis and Information Visualization Approach for Instant Messaging Applications**

*Authors: Shahnaz Pirzada, Nurul Hidayah Ab Rahman, Niken Dwi Wahyu Cahyani, Muhammad Fakri Othman*

**PAGE 237 – 246**

**Paper 30: Driving Maneuvers Recognition and Classification Using A Hybrid Pattern Matching and Machine Learning**

*Authors: Munaf Salim Najim Al-Din*

**PAGE 247 – 256**

**Paper 31: An Approach to Automatic Garbage Detection Framework Designing using CNN**

*Authors: Akhilesh Kumar Sharma, Antima Jain, Deevesh Chaudhary, Shamik Tiwari, Hairulnizam Mahdin, Zirawani Baharum, Shazlyn Milleana Shaharudin, Ruhaila Maskat, Mohammad Syafwan Arshad*

**PAGE 257 – 262**

**Paper 32: Tamper Proof Air Quality Management System using Blockchain**

*Authors: Vaneeta M, Deepa S R, Sangeetha V, Kamalakshi Naganna, Kruthika S Vasisht, Ashwini J, Nikitha M, Srividya H. R*

**PAGE 263 – 271**

**Paper 33: Optimized Strategy for Inter-Service Communication in Microservices**

*Authors: Sidath Weerasinghe, Indika Perera*

**PAGE 272 – 279**

**Paper 34: Deep Learning based Analysis of MRI Images for Brain Tumor Diagnosis**

*Authors: Srinivasarao Gajula, V. Rajesh*

**PAGE 280 – 286**

**Paper 35: Classification of Psychological Disorders by Feature Ranking and Fusion using Gradient Boosting**

*Authors: Saba Tahseen, Ajit Danti*

**PAGE 287 – 294**

**Paper 36: Public Response to the Legalization of The Criminal Code Bill with Twitter Data Sentiment Analysis**

*Authors: Deny Irawan, Dana Indra Sensuse, Prasetyo Adi Wibowo Putro, Aji Prasetyo*

**PAGE 295 – 303**

**Paper 37: Evaluation of QoS over IEEE 802.11 Wireless Network in the Implementation of Internet Protocols Mobility Supporting**

*Authors: Narimane Elhilali, Mostapha Badri, Mouncef Filali Bouami*

**PAGE 304 – 310**

**Paper 38: Using Deep Learning Algorithms to Diagnose Coronavirus Disease (COVID-19)**

*Authors: Nfayel Alanazi, Yasser Kotb*

**PAGE 311 – 320**

**Paper 39: Enhanced Optimized Classification Model of Chronic Kidney Disease**

*Authors: Shahinda Elkholy, Amira Rezk, Ahmed Abo El Fetoh Saleh*

**PAGE 321 – 331**

**Paper 40: Automated Categorization of Research Papers with MONO Supervised Term Weighting in RECAApp**

*Authors: Ivic Jan A. Biol, Rhey Marc A. Depositario, Glenn Geo T. Noangay, Julian Michael F. Melchor, Christopher C. Abalorio, James Cloyd M. Bustillo*

**PAGE 332 – 339**

**Paper 41: R-Diffset vs. IR-Diffset: Comparison Analysis in Dense and Sparse Data**

*Authors: Julaily Aida Jusoh, Sharifah Zulaikha Tengku Hassan, Wan Aezwani Wan Abu Bakar, Syarilla Iryani Ahmad Saany, Mohd Khalid Awang, Norlina Udin @ Kamaruddin*

**PAGE 340 – 346**

**Paper 42: A Fully Immersive Virtual Reality Cycling Training (vProCycle) and its Findings**

*Authors: Imran Bin Mahalil, Azmi Bin Mohd Yusof, Nazrita Binti Ibrahim, Eze Manzura Binti Mohd Mahidin, Ng Hui Hwa*

**PAGE 347 – 353**

**Paper 43: First Responders Space Subdivision Framework for Indoor Navigation**

*Authors: Asep Id Hadiana, Safiza Suhana Kamal Baharin, Zahriah Othman*

**PAGE 354 – 363**

**Paper 44: Leaf Diseases Identification and Classification of Self-Collected Dataset on Groundnut Crop using Progressive Convolutional Neural Network (PGCNN)**

*Authors: Anna Anbumozhi, Shanthini A*

**PAGE 364 – 373**

**Paper 45: Enhancing Image for CNN-based Diagnostic of Pediatric Pneumonia through Chest Radiographs**

*Authors: Vaishali Arya, Tapas Kumar*

**PAGE 374 – 380**

**Paper 46: Long Short-Term Memory for Non-Factoid Answer Selection in Indonesian Question Answering System for Health Information**

*Authors: Retno Kusumaningrum, Alfi F. Hanifah, Khadijah Khadijah, Sukmawati N. Endah, Priyo S. Sasongko*

**PAGE 381 – 388**

**Paper 47: Assessment of the Healthcare Administration of Senior Citizens from Survey Data using Sentiment Analysis**

*Authors: Ramona Michelle M. Magtangob, Thelma D. Palaoag*

**PAGE 389 – 394**

**Paper 48: Hierarchical Pretrained Deep Learning Features for the Breast Cancer Classification**

*Authors: Abeer S. Alsheddi*

**PAGE 395 – 402**

**Paper 49: A Survey on Attention-Based Models for Image Captioning**

*Authors: Asmaa A. E. Osman, Mohamed A. Wahby Shalaby, Mona M. Soliman, Khaled M. Elsayed*

**PAGE 403 – 412**

**Paper 50: Towards an Automatic Speech-to-Text Transcription System: Amazigh Language**

*Authors: Ahmed Ouhnini, Brahim Aksasse, Mohammed Ouanan*

**PAGE 413 – 418**

**Paper 51: Graphical user Interfaces Generation from BPMN (Business Process Model and Notation) via IFML (Interaction Flow Modeling Language) up to PSM (Platform Specific Model) Level**

*Authors: Abir Sajji, Yassine Rhazali, Youssef Hadi*

**PAGE 419 – 431**

**Paper 52: A Fuzzy Logic based Solution for Network Traffic Problems in Migrating Parallel Crawlers**

*Authors: Mohammed Faizan Farooqui, Mohammad Muqem, Sultan Ahmad, Jabeen Nazeer, Hikmat A. M. Abdeljaber*

**PAGE 432 – 441**

**Paper 53: A Privacy-Centered Protocol for Enhancing Security and Authentication of Academic Certificates**

*Authors: Omar S. Saleh, Osman Ghazali, Norbik Bashah Idris*

**PAGE 442 – 452**

**Paper 54: A Systematic Literature Review on AI Algorithms and Techniques Adopted by e-Learning Platforms for Psychological and Emotional States**

*Authors: Lubna A. Alharbi*

**PAGE 453 – 464**

**Paper 55: Privacy and Integrity Verification Model with Decentralized ABE and Double Encryption Storage Scheme**

*Authors: Amrutha Muralidharan Nair, R Santhosh*

**PAGE 465 – 471**

**Paper 56: Design of a Hybrid Recommendation Algorithm based on Multi-objective Collaborative Filtering for Massive Cloud Data**

*Authors: Xiaoli Zhou*

**PAGE 472 – 481**

Paper 57: Equally Spread Current Execution Load Modelling with Optimize Response Time Brokerage Policy for Cloud Computing

*Authors: Anisah Hamimi Zamri, Nor Syazwani Mohd Pakhrudin, Shuria Saaidin, Murizah Kassim*

**PAGE 482 – 491**

Paper 58: Research on the Derivative Rule and Estimation Methods of Intelligent High-Speed Railway Investment Estimation

*Authors: Yang Meng, Chuncheng Meng, Xiaochen Duan*

**PAGE 492 – 501**

Paper 59: Early Warning for Sugarcane Growth using Phenology-Based Remote Sensing by Region

*Authors: Sudianto Sudianto, Yeni Herdiyeni, Lilik Budi Prasetyo*

**PAGE 502 – 510**

Paper 60: Fault Tolerance Smart Egg Incubation System with Computer Vision

*Authors: Emiliyan Petkov, Teodor Kalushkov, Donika Valcheva, Georgi Shipkovenski*

**PAGE 511 – 517**

Paper 61: A Novel Hybrid Deep Learning Framework for Detection and Categorization of Brain Tumor from Magnetic Resonance Images

*Authors: Yousef Methkal Abd Algani, B. Nageswara Rao, Chamandeep Kaur, B. Ashreetha, K. V. Daya Sagar, Yousef A. Baker El-Ebiary*

**PAGE 518 – 527**

Paper 62: Multi Feature DCR based Drug Compound Selection and Recommendation System for Efficient Decision-Making using Genetic Algorithm

*Authors: ST. Aarthy, J. L. Mazher Iqbal*

**PAGE 528 – 533**

Paper 63: The Predictors of Mobile Banking Usage: A Systematic Literature Review

*Authors: Mohammed Abd Al-Munaf Hashim, Zainuddin Bin Hassan*

**PAGE 534 – 540**

Paper 64: A Novel Approach: Tokenization Framework based on Sentence Structure in Indonesian Language

*Authors: Johannes Petrus, Ermatita, Sukemi, Erwin*

**PAGE 541 – 549**

Paper 65: An Efficient Real-Time Weed Detection Technique using YOLOv7

*Authors: Ch. Lakshmi Narayana, Kondapalli Venkata Ramana*

**PAGE 550 – 556**

Paper 66: Sobel Edge Detection Algorithm with Adaptive Threshold based on Improved Genetic Algorithm for Image Processing

*Authors: Weibin Kong, Jianzhao Chen, Yubin Song, Zhongqing Fang, Xiaofang Yang, Hongyan Zhang*

**PAGE 557 – 562**

Paper 67: Compiler Optimization Prediction with New Self-Improved Optimization Model

*Authors: Chaitali Shewale, Sagar B. Shinde, Yogesh B. Gurav, Rupesh J. Partil, Sandeep U. Kadam*

**PAGE 563 – 571**

Paper 68: Design of an English Web-based Teaching Resource Sharing Platform based on Mobile Web Technology

Authors: Yan Zhang

PAGE 572 – 580

Paper 69: A Study of Encryption for Multimedia Digital Audio Security

Authors: Xiaodong Zhou, Chao Wei, Xiaotang Shao

PAGE 581 – 586

Paper 70: Digital Twins for Smart Home Gadget Threat Prediction using Deep Convolution Neural Network

Authors: Valluri Padmapriya, Muktevi Srivenkatesh

PAGE 587 – 594

Paper 71: A New Privacy-Preserving Protocol for Academic Certificates on Hyperledger Fabric

Authors: Omar S. Saleh, Osman Ghazali, Norbik Bashah Idris

PAGE 595 – 609

Paper 72: Breast Cancer Prediction using Machine Learning Models

Authors: Orlando Iparraguirre-Villanueva, Andrés Epifanía-Huerta, Carmen Torres-Cecién, John Ruiz-Alvarado, Michael Cabanillas-Carbonell

PAGE 610 – 620

Paper 73: Placement of Edge Servers in Mobile Cloud Computing using Artificial Bee Colony Algorithm

Authors: Bing Zhou, Bei Lu, Zhigang Zhang

PAGE 621 – 637

Paper 74: Erythematous-Squamous Disease Detection using Best Optimized Estimators of ANN

Authors: Rajashekar Deva, G .Narsimha

PAGE 638 – 647

Paper 75: Realizing the Quantum Relative Entropy of Two Noisy States using the Hudson-Parthasarathy Equations

Authors: Bhaveshkumar B. Prajapati, Nirbhay Kumar Chaubey

PAGE 648 – 655

Paper 76: Research on Automatic Detection Algorithm for Pedestrians on the Road Based on Image Processing Method

Authors: Qing Zhang

PAGE 656 – 661

Paper 77: Enhanced Multi-Verse Optimizer (TMVO) and Applying it in Test Data Generation for Path Testing

Authors: Mohammad Hashem Ryalat, Hussam N. Fakhouri, Jamal Zraqou, Faten Hamad, Mamon S. Alzboun, Ahmad K. Al hwaitat

PAGE 662 – 673

Paper 78: EFASFMM: A Unique Approach for Early Prediction of Type II Diabetics using Fire Fly and Semi-supervised Min-Max Algorithm

Authors: B. Manikyala Rao, Mohammed Ali Hussain

PAGE 674 – 681

Paper 79: Development of a Mobile Application to Reduce the Rate of People with Text Neck Syndrome

Authors: Rosa Perez-Siguas, Hernan Matta-Solis, Eduardo Matta-Solis, Hernan Matta-Perez, Luis Perez-Siguas, Randall Seminario Unzueta, Victoria Tacas-Yarcuri

PAGE 682 – 688

**Paper 80: An Early Warning Model for Intelligent Operation of Power Engineering based on Kalman Filter Algorithm**

*Authors: Haopeng Shi, Xiang Li, Pei Sun, Najuan Jia, Qiyang Dou*

**PAGE 689 – 698**

**Paper 81: Automated Pneumonia Diagnosis using a 2D Deep Convolutional Neural Network with Chest X-Ray Images**

*Authors: Kamila Kassylkassova, Batyrkhan Omarov, Gulnur Kazbekova, Zhadra Kozhamkulova, Mukhit Maikotov, Zhanar Bidakhmet*

**PAGE 699 – 708**

**Paper 82: Classification of Human Sperms using ResNet-50 Deep Neural Network**

*Authors: Ahmad Abdelaziz Mashaal, Mohamed A. A. Eldosoky, Lamia Nabil Mahdy, Kadry Ali Ezzat*

**PAGE 709 – 713**

**Paper 83: Research on Image Sharpness Enhancement Technology based on Depth Learning**

*Authors: Wenbao Lan, Chang Che*

**PAGE 714 – 722**

**Paper 84: Fall Detection and Monitoring using Machine Learning: A Comparative Study**

*Authors: Shaima R. M Edeib, Rudzidatul Akmam Dziyauddin, Nur Izdihar Muhd Amir*

**PAGE 723 – 728**

**Paper 85: Online Teaching Design and Evaluation of Innovation and Entrepreneurship Courses in the Context of Education Internationalization**

*Authors: Chengshe Xing*

**PAGE 729 – 738**

**Paper 86: Investigating Group Distributionally Robust Optimization for Deep Imbalanced Learning: A Case Study of Binary Tabular Data Classification**

*Authors: Ismail. B. Mustapha, Shafaatunnur Hasan, Hatem S Y Nabbus, Mohamed Mostafa Ali Montaser, Sunday Olusanya Olatunji, Siti Maryam Shamsuddin*

**PAGE 739 – 748**

**Paper 87: A Biologically Inspired Appearance Modeling and Sample Feature-based Approach for Visual Target Tracking in Aerial Images**

*Authors: Lili Pei, Xiaohui Zhang*

**PAGE 749 – 760**

**Paper 88: A Study on Distance Personalized English Teaching Based on Deep Directed Graph Knowledge Tracking Model**

*Authors: Lianmei Deng*

**PAGE 761 – 770**

**Paper 89: A Visual Target Representation using Saliency Detection Approach**

*Authors: Shekun Tong, Chunmeng Lu*

**PAGE 771 – 780**

**Paper 90: Building a Machine Learning Powered Chatbot for KSU Blackboard Users**

*Authors: Qubayl Alqahtani, Omer Alrwais*

**PAGE 781 – 789**

Paper 91: WEB-based Collaborative Platform for College English Teaching

*Authors: Yuwan Zhang*

PAGE 790 – 800

Paper 92: Predictions of Cybersecurity Experts on Future Cyber-Attacks and Related Cybersecurity Measures

*Authors: Ahmad Mtair AL-Hawamleh*

PAGE 801 – 809

Paper 93: BERT Model-based Natural Language to NoSQL Query Conversion using Deep Learning Approach

*Authors: Kazi Mojammel Hossen, Mohammed Nasir Uddin, Minhazul Arefin, Md Ashraf Uddin*

PAGE 810 – 821

Paper 94: Privacy-Preserving and Trustless Verifiable Fairness Audit of Machine Learning Models

*Authors: Gui Tang, Wuzheng Tan, Mei Cai*

PAGE 822 – 832

Paper 95: An OCR Engine for Printed Receipt Images using Deep Learning Techniques

*Authors: Cagri Sayallar, Ahmet Sayar, Nurcan Babalik*

PAGE 833 – 840

Paper 96: A Survey on Blockchain Technology Concepts, Applications and Security

*Authors: Asma Mubark Alqahtani, Abdulmohsen Algarni*

PAGE 841 – 847

Paper 97: An Autonomous Role and Consideration of Electronic Health Systems with Access Control in Developed Countries: A Review

*Authors: Mohd Rafiz Salji, Nur Izura Udzir*

PAGE 848 – 860

Paper 98: Semi-supervised Method to Detect Fraudulent Transactions and Identify Fraud Types while Minimizing Mounting Costs

*Authors: Chergui Hamza, Abrouk Lylia, Cullot Nadine, Cabioch Nicolas*

PAGE 861 – 870

Paper 99: Liver Disease Prediction and Classification using Machine Learning Techniques

*Authors: Srilatha Tokala, Koduru Hajarathaiah, Sai Ram Praneeth Gunda, Srinivasrao Botla, Lakshmikanth Nalluri, Pathipati Nagamanohar, Satish Anamalamudi, Murali Krishna Enduri*

PAGE 871 – 878

Paper 100: Image Super-Resolution using Generative Adversarial Networks with EfficientNetV2

*Authors: Saleh AlTakroui, Norliza Mohd Noor, Norulhusna Ahmad, Taghreed Justinia, Sahnus Usman*

PAGE 879 – 887

Paper 101: A Transformer Seq2Seq Model with Fast Fourier Transform Layers for Rephrasing and Simplifying Complex Arabic Text

*Authors: Abdullah Alshaqiti, Ahmad Alkhodre, Abdallah Namoun, Sami Albouq, Emad Nabil*

PAGE 888 – 898

Paper 102: Unsupervised Feature Learning Methodology for Tree based Classifier and SVM to Classify Encrypted Traffic

*Authors: RAMRAJ S, Usha G*

PAGE 899 – 904

**Paper 103: Indoor Pollutant Classification Modeling using Relevant Sensors under Thermodynamic Conditions with Multilayer Perceptron Hyperparameter Tuning**

*Authors: Percival J. Forcadilla*

**PAGE 905 – 916**

# Organizational Digital Transformations and the Importance of Assessing Theoretical Frameworks such as TAM, TTF, and UTAUT: A Review

Bibhu Dash<sup>1</sup>, Pawankumar Sharma<sup>2</sup>, Swati Swayamsiddha<sup>3</sup>

School of Computer and Information Sciences, University of the Cumberland, Williamsburg, KY USA<sup>1,2</sup>  
School of Electronics Engineering, KIIT University, Bhubaneswar, India<sup>3</sup>

**Abstract**—In this era of Industry 5.0, businesses worldwide are attempting to gain competitive advantages, increase profits, and improve consumer engagement. To achieve their goals, all businesses undergo extensive digital transformations (DT) by implementing cutting-edge technologies such as cloud computing, artificial intelligence (AI), the Internet of Things, and blockchain, among others. DT is a costly journey, including strategy, people, and technology. At the same time, many digitization efforts are failing miserably, resulting in project abandonment, loss of critical stakeholder trust, and the dismissal of important staff. Poor strategy, which may have pre-evaluated organizational flexibility and cultural misfits, is often criticized. As a result, it is critical to extensively investigate theoretical frameworks such as the Technology Acceptance Model (TAM), Task Technology Fit (TTF), and Unified Theory of Acceptance and Use of Technology (UTAUT), which were developed via significant research into various organizational kinds. All of these aspects are covered in this work by evaluating academic papers from the IEEE, Scopus, and Web of Science databases and reaching conclusions in future sections.

**Keywords**—Data growth; digital transformations; TAM; TTF; UTAUT; sustainability; FTM

## I. INTRODUCTION

With the increasing rise of social media, proper data storage and retrieval in a modern data-driven company model are important for long-term viability and competitive advantage. In the ever-expanding business sector, AI and machine learning are emerging as feasible digital domains for information storage and recovery, promising to improve access to knowledge and effective decision-making [1]. A recent Google analysis shows DT themes and trends rapidly rising, the most contested subjects worldwide [2]. Digitalization provides greater functional assistance to clients while also illuminating performance and producing more substantial revenue streams [3]. The following factors have been recognized as driving the growth of unstructured big data and digitalization worldwide.

### A. Data Growth

As data storage costs continue to decline due to the emergence of the cloud, organizations retain substantial volumes of transactional data for analysis and research [4]. Every industry type is seeing an increase in data due to the daily growth of e-transaction volumes in institutions. The dollar equivalent of electronic transactions will likely total

roughly \$118.3 billion globally in 2021, according to a recent Business Wire estimate [5], with developing markets increasing 15–20% faster than developed ones [6].

For a few decades, the objectives of businesses and the adoption of digitalization have changed due to the tremendous growth in data volume and quality. Because of these enormous data quantities, governance and regulatory organizations are under pressure to manage and preserve sensitive data. Data growth over the years depicts the price development for storing a gigabit of data from 1966 to 2020, as shown below. The cost of storing a gigabyte of data decreased drastically from \$1.05 million in 1966 to \$0.02 in 2017 [7]. Financial institutions are prompted by the cheaper storage costs to retain and process this enormous volume of data for important insight retrieval, organizational development, and decision-making [4].

### B. Swing in the Business Model

The convenience and cost of online transactions are made possible by digitalizing essential products and services. Because of the aforementioned considerations and the developing global market, global institutions operate differently than they did a few decades ago. Self-service analytics and unstructured data storage are helping modern digital enterprises hold onto market share in the face of escalating intra- and inter-domain competition. With the evolution of smartphones, Know Your Customer (KYC), and Know Your Product (KYP), the pressure on businesses mounted to promote speed, efficacy, and quality by going digitalization [7]. Thus, in contemporary customer-focused business models, data security, storage, and meaning extraction are gaining center stage.

## II. DIGITALIZATION AND ITS CHALLENGES

Data is the new gold as digitization approaches its zenith [8]. As data growth happens with constantly decreasing storage costs, organizations get the much-required push to be more data-driven than before in daily operations, as shown in Fig. 1. This study was significant because it expanded our understanding of the factors that push firms to adopt a data-driven approach to all facets of daily operations and works as catalyst for organizational sustainability. But making big data useful and facilitating faster information retrieval is now enterprises' key issue [9]. According to the study mentioned above by Chowdhary [9], firms can examine consumer insights and behavioral trends but cannot take particular data-driven

action because of decision management issues. This failure can result from issues with organizational implementation or poor C-suite execution. Some businesses are also aware of the value of data and digitization. However, the difficulty they have in filling the positions is the absence of human resources with the necessary degree of expertise and strategy. Below are some discussions of why an organization's digitalization journey fails and the cause of this.

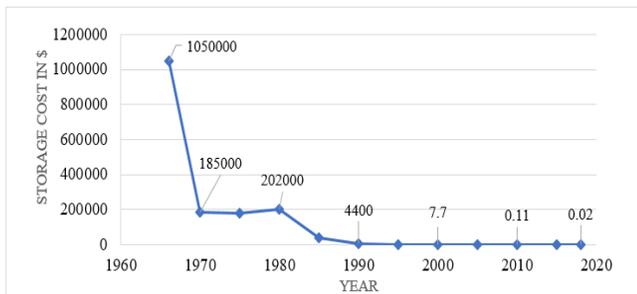


Fig. 1. Cost per gigabyte of data storage over time

#### A. Lack of Transformation Objectives

Without a solid leadership vision or goal, achieving digitalization objectives is very hard. The fact that digital transformation never had a chance to succeed is one of the key reasons it failed [6, 7]. It is generally viewed as a publicity stunt that will give a company an air of optimism, imagination, and the future. Therefore, no clear objectives have been set to guide digital transformation efforts, which are typically solutions looking for a problem to solve. This leads to a company releasing the latest technology to release emerging innovations without having performance indicators in place. When the project is closely examined, it dragged out, exceeded budget, and ultimately failed.

#### B. No Planning for Change Management

Agile operation and change management are essential in the modern digitalization journey. Businesses may decide how their daily operations will run while introducing new technology with the help of a change management strategy [10]. Whether it's a new ML model, Cloud adoption, ERP, or a different data pipeline build, a change management plan outlines how operations will run until the migration is successfully completed. If change management is properly outlined, the project is definitely going to lose its control, both concerning cost and time, and it is bound to fail.

#### C. Organizational Culture and Inner Endurance to Change

Recent research shows that 40–50% of Executives believe their teams never utilize the majority of the functionality offered by digitalized products because of internal resistance to change [10]. This problem affects more than just product development, sales, or finance; it also affects disciplines like marketing, business development, and architecture design. Professionals typically think that their own systems they've created are more dependable than other software. Similarly, they could not have the training required to properly use a new product and opt to put it on hold for the time being. It is majorly a cultural thing affecting many companies globally and ultimately leads to digital transformation failure.

#### D. Too Fast in Bringing Transformation through Technology

To be the first to market (FTM), many firms set an accelerated schedule for a digital transformation program shorter than their implementation partners and consultants advise. As a result, because the leadership needs to be made aware of the technical effort involved in transferring to alternative IT systems, some of the critical processes or Key Performance Indicators (KPIs) required to make the switch are skimmed on for enterprise adaptability and sustainability[10]. It leads to performance and quality compromise and ultimately leads to failure.

#### E. Poor Vision and Adoption of new Technology

Different corporate departments have various digitalization aspirations. For instance, marketing may desire to boost traffic regardless of whether it converts to customers, the engineering department may want to solve challenges that will improve performance, and the human resource team may want to meet hiring goals while not engaging with the rest of the company often [7, 10]. It could be challenging to get funding, get departmental buy-in, or commit to doing things to make a digital transformation program successful without a common organizational vision.

When only some people utilize the new IT solution as regularly as the leadership had planned, a company may effectively move to it just to run into a new challenge. When they are uncertain about the training they have received, employees understandably shy away from new technologies. Similarly, there is always a time when employees are reluctant to try out new tools in favor of tried-and-true methods. Organizations must create a strategy to enhance the digital adoption of new apps to achieve project Return on Investment (ROI). Sometimes it happens that junior staff members in an organization see the need for an improved alternative to the one they now use. They ask the leadership for a new response as a result. They are frequently declined, and even if they are grudgingly approved, the project does not obtain the required cash, help, and operational support to be effective throughout the whole organization. Sometimes these forced buy-ins adversely impact employee commitment and break the trust in the digitalization journey before it really starts [10].

#### F. Lack of Basics and Ignoring Customer Expectations

A digital transformation initiative should prioritize enhancing the company's capacity to serve its key customers. Any digital transformation effort will only succeed if it enables us to give value to our customers, regardless of how much better the design is, how quickly you put it out, or how good our implementation partners were. If we ignore our core loyal customers and design a new digital app or product that lacks our domain-specific information, then it's a waste of time and money, which will not meet business needs [8].

The biggest enterprises are leaders in technology, product, and innovation, but they also need a better history of losing millions to a failed digitalization journey [8, 9]. These businesses made unavoidable mistakes when using digitalization to empower their operations. The table below (see Table I) shows some of these world-famous brands and

the gaps in their journey to lose financially during their digital adoptions [10].

TABLE I. INFAMOUS DIGITALIZATION FAILURES AND THEIR REASONS

Organization	Targeted for	Reason for Failure
Hershey (1996)	Hershey was targeting more powerful ERP systems to replace their legacy IT systems.	1- No clear vision and trying to execute a pet project. 2-The transformation was rushed without proper testing.
HP (2003)	HP planned to stop legacy ERP systems and move to SAP for better customer and sales tracking.	1- Migration exceeded the target time frame. 2- Bad change management planning. 3- Didn't plan for a backup if this ERP implementation fails.
Miller Coors (2013)	MC planned digitalization through new app development and advanced ERP implementation to boost the supply chain.	The project went 3 years without any end with bad vendor selection and bad planning, ending with a lawsuit.
Revlon(2018)	Revlon was targeting to migrate its new ERP solutions to 22 countries where they operate	Due to bad planning, poor change management, and implementation issues, the company lost \$64 million, and the market cap dropped by 6.9%.

### III. THEORETICAL FRAMEWORKS AND THEIR ASSESSMENT

Academics develop theoretical frameworks through in-depth research, surveys, and testing. Theoretical frameworks aid in understanding the implicit theory in adopting, developing, or clarifying expressly stated norms before implementing new initiatives, strategies, and regulations. It aids in reducing biases that might skew our interpretations in a novel environment owing to cultural and technological incompatibilities in an organizational or public context. One theoretical framework may be used to analyze other theoretical frameworks, which can change our perspective and reduce the risks of theory selection and application [11]. Theoretical frameworks help us recognize the boundaries of our study's scope by helping us understand the nature of our research problem. Better and clearer theoretical framework analysis helps us improve our decisions, develop better strategies, and understand our objectives.

Digital transformations are expensive, people-focused, and strategy-driven. It is, therefore, quite challenging to select the ideal theoretical framework to analyze the benefits of a digitalization project. Quantitative program or project design and theoretical underpinnings are interrelated. The study's goals and a comprehensive literature review are used to select the research design. Quantitative project planning employs deductive reasoning, which starts with choosing the theoretical framework that will provide the project with a solid foundation and direction. The early sections of a quantitative research proposal include theoretical frameworks to provide the justification for the inquiry. To direct the methods we use, choosing the right theoretical framework is essential. Thus, the study of the right methodology will provide conclusions that

are compatible with the organization, its end goal, and its culture. The details about the theoretical framework needed to study before adopting digital transformations are discussed below in detail.

#### A. TAM

Decision-making about digital transformation and organizational sustainability can be enhanced using Davis' [11] Technological Acceptance Model. The study's guiding principle is usefulness and usability, which served as the foundation for the TAM model. TAM was first designed to provide a logical framework for evaluating the user acceptability of a certain information system or piece of information technology [12, 13]. TAM is developed and utilized regularly in fields other than IT, such as healthcare, retail, and finance. Perceived utility (PU) and perceived ease of use (PEU) are the two main components of TAM (see Fig. 2). (PEU). These criteria were all created to characterize the usability and effectiveness of new procedures or technologies. External factors impact both PU and PEU, favorably affecting users' feelings about using the target system. Additionally, the variables affect users' use behavior (UB) on the target system and process [13]. Fig. 2 shows how individuals and organizations feel about using the TAM model to analyze unstructured data and gain insights for better outcomes.

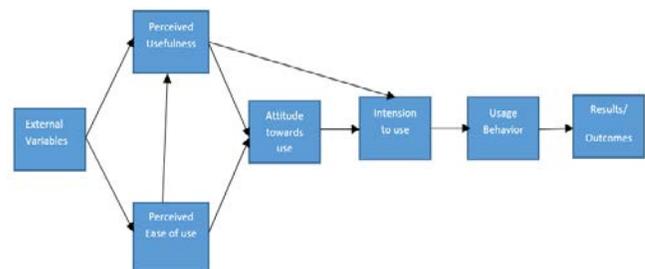


Fig. 2. Technology acceptance model and key attributes

According to contingency theory, the compatibility of an organization and its environment is a necessary condition for organizational performance. Verma et al. [14] investigated the adoption of self-serve analytics and digitalization by utilizing the TAM model to examine how system features affected managers' opinions. Despite the past effort, research has yet to employ a complete strategy to experimentally evaluate the technical fit, organizational fit, and environmental fit views. To evaluate the influence of technology fit, organization fit, and environment fit on applying data analysis, KPIs, and insights for data-driven performance outcomes, the TAM and TTF models were merged in this study.

#### B. TTF

More effectively than technical, environmental, and individual factors, Task Technology Fit (TTF) is a framework that encourages innovation and adaptability [14]. Adopting cutting-edge technology by a business is influenced by various elements, including financial competence, organizational growth, technical services, and external environmental factors [15]. The primary external factors influencing adoption were a relative advantage, competitive pressure, and government support. Individuals' behaviors in converting inputs into outputs were referred to as task characteristics [16]. This study

aimed to predict performance using a combination of organizational structure, technology utilization, and the application of information technology to strategy better to account for the unpredictability of the external environment. The task-technology fit hypothesis was developed in the adoption of information systems, and the use of the idea of fit to assess a technology's influence on business performance has increasingly increased [16]. Generally speaking, this idea contends that the effectiveness of an information system depends on how effectively a job and technology interlock (see Fig. 3).

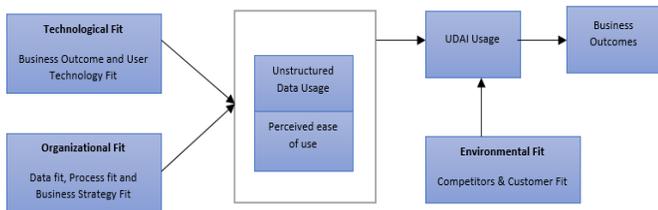


Fig. 3. TTF model for digital adoption

From the TTF perspective, novelty, output quality, and compatibility are crucial requirements for effective knowledge use and better outcomes. According to Wu and Chen [17], it still needs to be determined which elements, based on the TTF model, affect the performance of technology-driven companies. According to Wu and Chen, increasing TTF from the perspectives of technical, organizational, and environmental fit may impact technology-driven performance in any corporation to close the gaps. The Fintech mechanism is driven by business, and technological advancements are essential to any current Fintech perspective [18]. Excellent customer service and the development of distinctive products are essential for a firm to succeed in the Fintech industry.

### C. UTAUT

The TAM model is the foundation for the UTAUT, which is enhanced by "adding social influence and positive factors" [19]. A technological acceptance model using UTAUT was created by Venkatesh et al. [20] and describes users' intentions to adopt information technology, digital transformation, and subsequent user behavior. The theory strongly emphasizes four key concepts: (a) performance expectancy, (b) effort expectancy, (c) social characteristics, and (d) enabling conditions. This section's first three components cover usage behavior, while its final element emphasizes user behavior (see Fig. 4). UTAUT, sometimes referred to as child TAM, accounts for 50% of the variance in actual usage or user behavior (UB) and 70% of the variance in behavioral intention (BI) [21, 22].

A study of customer behavior is essential in this situation because it clarifies studies that show that customers' perceptions of performance improvement from technology are based on performance expectations [20]. TTF is used for utilities, UTAUT is used for companies, and the TAM framework is utilized for persons. The main focus of the new UTAUT model is on the economic or social effects of technology use. Fig. 4 shows how data-driven finance has changed from Fintech to TechFin [18-23] due to the major

changes in the worldwide regulatory environment in the finance industry following the 2008 financial crisis.

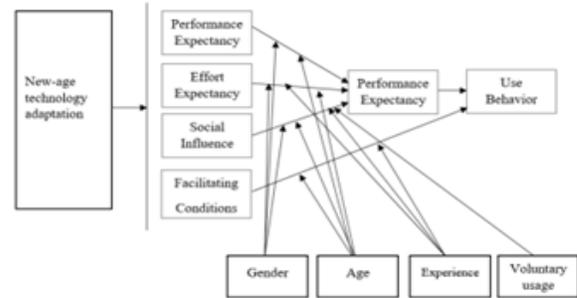


Fig. 4. Organizational technology adaptation applying the UTAUT model [18]

Decentralized norms and regulations provide AI, IoT, and modern technologies a platform to participate in money absorption, securities dealing, and illegal fund-raising that goes beyond what applies to it [20]. This study's concept and framework provide a new business and technical innovation paradigm. The study's findings provide theoretical underpinnings and practical advice for further education. Every corporation's technology and resources are its true assets [24], and business goals, procedures, and technical innovation determine the results.

### IV. KEY TAKEAWAYS FROM TAM, TTF, AND UTAUT

- Both internal and external elements are included in the TAM model. To ascertain if a new endeavor or computer project will be embraced by its potential users, TAM places a high value on "perceived utility" and "perceived ease of use." The TAM model examines exterior elements, including utility, content, pricing, and design, to examine their links with perceived usefulness and perceived usability [16]. Before we continue to investigate the idea of reasoned action, these external elements that impact organizational culture or commercial acceptance must be addressed (TRA) [25]. The TAM has a weakness, though, in that social influence is disregarded and has less of an impact in corporate settings today because it is based on individual ideas.
- In the current research, TTF is determined by evaluating how well the system function satisfies the demands of each specific activity [15]. Both corporate settings and specific workers can benefit from TTF. Every action we take inside the business must be assessed from both corporate and personal viewpoints to ensure that it meets the needs of all stakeholders. TTF is the connection between a task's needs, a person's aptitudes, and a digital device's and its software's capability. Additionally, TTF has been connected to the standard of personal performance, which may be applied in a wider framework of thinking about how information technology affects human performance [26]. The clear message here is that anything incompatible with people cannot be useful in an organizational setting.

- The UTAUT model is crucial because it compares the most popular technology acceptance theories and offers empirical insights into how people embrace new technologies. An important aspect of this paradigm that affects whether people or organizations should utilize the new system is the effect of social or competitive factors [15, 26]. A low or high social influence score may impact a company's journey toward digital transformation. This study describes how social impact influences executives' and workers' motivation, ultimately affecting organizational decision-making. Fig. 4 loudly highlights the impact of social influence and organizational cultural change as an output of that (see Fig. 5, which explains that the journey of necessity starts with a journey of influence).
- The veracity and viability of theoretical models for digitalization validation increase along with technological advancement. The interactivity, adaptability, and brilliance of digital systems are now considered prior elements in the UTAUT model. All three above models showed that attitude was fundamental to behavioral intentions and usage behaviors, partially mediated the effects of exogenous constructs on behavioral intentions, and directly affected organizational usage behaviors. This was shown to be true in both direct and indirect ways. Numerous implications for theory and practice are made using the study findings, and conclusions are formed.

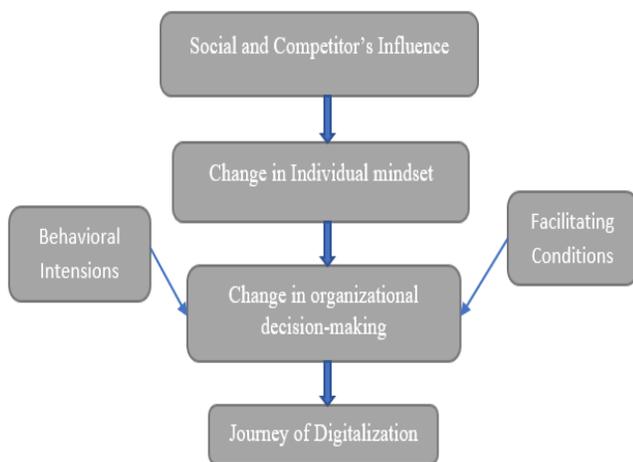


Fig. 5. UTAUT theory of social influence and digitization drive

## V. LIMITATIONS

The construction and analysis of theoretical frameworks aid in clarifying our implicit theory in a more specified manner. It is beneficial to evaluate other alternatives as well to avoid biases that may influence our understanding. However, it's different, and we need to realize that no one model that fits all. These frameworks are ancient and were created with a few scenarios in mind. However, contemporary organizations have complicated processes and needs. So, no single model can address all of their questions here. As a result, planners, project managers, and executives must investigate all three

frameworks and associated alternatives to prevent biases while attempting to match these models with some considerations and aspects to be avoided.

## VI. RECOMMENDATIONS

Because of the worldwide influences of 5G, smartphones, and shifting client sentiments, digital transformation is very necessary for modern enterprises [24]. Companies increasingly use people, technology, and procedure to their advantage in a multidimensional environment. But because it is expensive and cultural in character, affecting all aspects of the company domains, it is crucial to prepare it well. It is very much recommended to employ tested techniques and frameworks to analyze any positive and negative effects of any such major projects before beginning them. The existing organizational stance, its short- and long-term goals, and the cultural background may all be accessed using the theoretical frameworks TAM, TTF, and UTAUT [27, 28]. It is strongly advised that companies thoroughly research these works before selecting the right technology, vendor, and change management process to avoid misfiring in the future [29, 30]. Additionally, study results may affect behavior motivators and outside variables when analyzing executive-level intents and may serve as a basis for future analyses of the DT projects.

## VII. CONCLUSION

The practical assessment of theoretical frameworks, their drawbacks, the consequences for ongoing research, and sensible advice are highlighted in this study. This paper emphasizes the value of theoretical frameworks and explains why they are essential in post-pandemic organizations before beginning any new, expensive, and labor-intensive digitalization endeavor. Building software using these frameworks during pre-project strategy is crucial to achieving both qualitative and quantitative goals. As these models provide comprehensive perspectives of technology acceptability, technology use, and behavioral intents, future research is advised to improve data collecting from Fortune 500 businesses internationally and dialogues, including executives and stakeholders.

## ACKNOWLEDGMENT

Our chair and esteemed professor, Dr. Azad Ali, is acknowledged for his guidance and assessment of this work throughout our journey. Dr. Ali's inputs and suggestions are extremely beneficial in completing this systematic review.

## REFERENCES

- [1] Guo, H., & Polak, P. (2021). Artificial Intelligence and Financial Technology FinTech: How AI Is Being Used Under the Pandemic in 2020. In *The Fourth Industrial Revolution: Implementation of Artificial Intelligence for Growing Business Success* (pp. 169-186). Springer, Cham.
- [2] Chen, C. C., Huang, H. H., & Chen, H. H. (2020). NLP in FinTech applications: past, present, and future. *arXiv preprint arXiv:2005.01320*.
- [3] Chang, Y., Wong, S. F., Lee, H., & Jeong, S. P. (2016). What motivates Chinese consumers to adopt FinTech services: A regulatory focus theory. In *Proceedings of the 18th annual international conference on electronic commerce: e-commerce in smart connected world* (pp. 1-3).
- [4] Arslanian, H., & Fischer, F. (2019). The future of finance: The impact of FinTech, AI, and crypto on financial services. Springer.(Alberti, 2022) - 5

- [5] Koh, W. C., Kose, M. A., Nagle, P. S. O., Ohnsorge, F., & Sugawara, N. (2020). Debt and financial crises.
- [6] Mearian, L. (2017, March 23). CW@50: Data Storage goes from \$1m to 2 cents per gigabyte (+video). Computerworld. Retrieved May 8, 2022, from <https://www.computerworld.com/article/3182207/cw50-data-storage-goes-from-1m-to-2-cents-per-gigabyte.html>
- [7] Melville, N., Kraemer, K., & Gurbaxani, V. (2004). Review: Information Technology and Organizational Performance: An Integrative Model of IT Business Value. *MIS Quarterly*, 28(2), 283. <https://doi.org/10.2307/25148636>
- [8] Xu, H. Y. (2017). China's Internet Financial Risks and Countermeasures. In International Conference on Financial Management, Education and Social Science (FMES 2017).
- [9] Chowdhary, K. R. (2020). Natural Language Processing. *Fundamentals of Artificial Intelligence*, 603-649. [https://doi.org/10.1007/978-81-322-3972-7\\_19](https://doi.org/10.1007/978-81-322-3972-7_19)
- [10] Rohn, S. (2022, November 7). 4 high-profile digital transformation failures (+causes). The Whatfix Blog | Drive Digital Adoption. Retrieved January 21, 2023, from <https://whatfix.com/blog/digital-transformation-failures/>
- [11] Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, 13(3), 319-340. <https://doi.org/10.2307/249008>
- [12] Brock, V., & Khan, H. U. (2017). Big data analytics: does organizational factor matters impact technology acceptance?. *Journal of Big Data*, 4(1), 1-28.
- [13] Zhong, H., & Xiao, J. (2015, September). Apply technology acceptance model with big data analytics and unity game engine. In 2015 6th IEEE International Conference on Software Engineering and Service Science (ICSESS) (pp. 19-24). IEEE.
- [14] Verma, S., Bhattacharyya, S. S., & Kumar, S. (2018). An extension of the technology acceptance model in the big data analytics system implementation environment. *Information Processing & Management*, 54(5), 791-806. <https://doi.org/10.1016/j.ipm.2018.01.004>
- [15] Eze, S. C., Awa, H. O., Okoye, J. C., Emecheta, B. C., & Anazodo, R. O. (2013). Determinant factors of information communication technology (ICT) adoption by government-owned universities in Nigeria: A qualitative approach. *Journal of Enterprise Information Management*.
- [16] Gangwar, H. (2020). Big Data Analytics Usage and Business Performance: Integrating the Technology Acceptance Model (TAM) and Task Technology Fit (TTF) Model. *Electronic Journal of Information Systems Evaluation*, 23(1), pp45-64.
- [17] Wu, B., & Chen, X. (2017). Continuance intention to use MOOCs: Integrating the technology acceptance model (TAM) and task technology fit (TTF) model. *Computers in human behavior*, 67, 221-232. <https://doi.org/10.1016/j.chb.2016.10.028>
- [18] Legowo, M. B., Subanija, S., & Sorongan, F. A. (2020). Role of FinTech Mechanism to Technological Innovation: A Conceptual Framework. *International Journal of Innovative Science and Research Technology*, 5(5), 1-6.
- [19] Brown, S. A., Dennis, A. R., & Venkatesh, V. (2010). Predicting collaboration technology use: Integrating technology adoption and collaboration research. *Journal of Management Information Systems*, 27(2), 9-53.
- [20] Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User Acceptance of Information Technology: Toward a Unified View. *MIS Quarterly*, 27(3), 425. <https://doi.org/10.2307/30036540>
- [21] Jain, R., Garg, N., & Khera, S. N. (2022). Adoption of AI-Enabled Tools in Social Development Organizations in India: An Extension of UTAUT Model. *Frontiers in Psychology*, 13.
- [22] Hoekstra, R., Kiers, H., & Johnson, A. (2012). Are assumptions of well-known statistical techniques checked, and why (not)?. *Frontiers in Psychology*, 3, 137.
- [23] Melville, N., Kraemer, K., & Gurbaxani, V. (2004). Review: Information Technology and Organizational Performance: An Integrative Model of IT Business Value. *MIS Quarterly*, 28(2), 283. <https://doi.org/10.2307/25148636>
- [24] Hong, S. H., & Yu, J. H. (2018, September). Identification of external variables for the Technology Acceptance Model (TAM) in the assessment of BIM application for mobile devices. In IOP Conference series: materials science and engineering (Vol. 401, No. 1, p. 012027). IOP Publishing.
- [25] Dash, B., Sharma, P., & Ali, A. (2022). Federated Learning for Privacy-Preserving: A Review of PII Data Analysis in Fintech. *International Journal of Software Engineering & Applications (IJSEA)*, 13(4).
- [26] Fuller, R. M., & Dennis, A. R. (2009). Does fit matter? The impact of task-technology fit and appropriation on team performance in repeated tasks. *Information Systems Research*, 20(1), 2-17.
- [27] Dash, B., & Ansari, M. F. (2022). Self-service analytics for data-driven decision making during COVID-19 pandemic: An organization's best defense. *Academia Letters*, 2.
- [28] Kang, J., & Van Ouytsel, J. (2023). Are Investors Willing to Use Zoom for Entrepreneurs' Pitch Presentations?. *Information*, 14(2), 107.
- [29] [Yadav, U. S., Tripathi, R., & Tripathi, M. A. (2022). Effect of Digital and Financial Awareness of Household Womens on the Use of Fin-Tech in India: Observing the Relation with (Utaut) Model. *Journal of Sustainable Business and Economics*, 5(3), 18-26.
- [30] Bommer, W. H., Rana, S., & Milevoj, E. (2022). A meta-analysis of eWallet adoption using the UTAUT model. *International Journal of BankMarketing*.

# Adversarial Sampling for Fairness Testing in Deep Neural Network

Tosin Ige<sup>1</sup>, William Marfo<sup>2</sup>, Justin Tonkinson<sup>3</sup>, Sikiru Adewale<sup>4</sup>, Bolanle Hafiz Matti<sup>5</sup>

Department of Computer Science, University of Texas at El Paso, Texas, USA<sup>1,2,3</sup>

Department of Computer Science, Virginia Tech. SW Blacksburg, Virginia, USA<sup>4</sup>

Department of Mathematics and Statistics, Austin Peay State University, Tennessee, USA<sup>5</sup>

**Abstract**—In this research, we focus on the usage of adversarial sampling to test for the fairness in the prediction of deep neural network model across different classes of image in a given dataset. While several framework had been proposed to ensure robustness of machine learning model against adversarial attack, some of which includes adversarial training algorithm. There is still the pitfall that adversarial training algorithm tends to cause disparity in accuracy and robustness among different group. Our research is aimed at using adversarial sampling to test for fairness in the prediction of deep neural network model across different classes or categories of image in a given dataset. We successfully demonstrated a new method of ensuring fairness across various group of input in deep neural network classifier. We trained our neural network model on the original image, and without training our model on the perturbed or attacked image. When we feed the adversarial samplings to our model, it was able to predict the original category/ class of the image the adversarial sample belongs to. We also introduced and used the separation of concern concept from software engineering whereby there is an additional standalone filter layer that filters perturbed image by heavily removing the noise or attack before automatically passing it to the network for classification, we were able to have accuracy of 93.3%. Cifar-10 dataset have ten categories of dataset, and so, in order to account for fairness, we applied our hypothesis across each categories of dataset and were able to get a consistent result and accuracy.

**Keywords**—Adversarial machine learning, adversarial attack; adversarial defense; machine learning fairness; fairness testing; adversarial sampling; deep neural network

## I. INTRODUCTION

With some of the latest advances in artificial intelligence, deep learning (DL) can now be applied in areas as diverse as, face recognition system [19], fraud detection system [20], and natural language processing (NLP) [21]. As deep neural network model continues to be increasingly associated with important decision in our daily life, we cannot just view it as only a mathematical abstraction but also as a technical system for the modern society[22],[18][17]. Apart from looking at the various metrics to better understand and evaluate the logic behind the prediction of machine learning model, it is also imperative to look at the ethics in order not to infringe on people's privacy in which the ability to ensure fairness across all groups without bias in dnn model prediction is of serious concern to the community[23], while it is possible to have intentional or unintentional discriminatory pattern in dataset[24] which are being used to train a dnn model, it is imperative and to have some mechanism to identify such

discrimination in dataset before training model with it as such discrimination in dataset are eventually passed onto the trained model when the model is trained with discriminatory dataset especially when the discrimination is among the minority and the vulnerable in the society.

There are several forms in which discrimination can exist in dataset set, some of the forms includes group discrimination [25], [26] and individual discrimination [27], [28],[29]. Some of this discrimination can also be defined over a set of certain attributes such as gender, race age and so on., and when a machine learning model is trained with a discriminatory dataset, such discrimination in dataset is always passed to the trained model which make the model to make bias prediction over the same group being discriminated against in the dataset and this can be seen when ML model makes different decisions for different individuals (individual discrimination) or subgroups (group discrimination). Note that the set of protected attributes is often application-dependent and given in advance.

Our research work is focussed on the usage of adversarial sampling to test for the fairness in the prediction of deep neural network model across different classes of image in a given dataset. We are not dealing with the problem of individual discrimination or samples that differs only by some protected features. We aimed to use adversarial sampling to test for fairness in a dnn model, while also making an avenue for scaling the fairness through the misclassification rate across all group of image. Several adversarial samples were generated from the original image through several adversarial sampling techniques which includes Calini & Wagner, fast gradient sign method (FGSM), adversarial patch, gradient base evasion, and projected Gradient Descent (PGD). Although proposals and conceptual framework had been researched and formulated to address the issue of fairness in ML model [30], [31], [32]. One example is THEMIS randomly samples each attribute within its domain and identifies those discriminative samples [30], and also AEQUITAS which aims to improve the testing effectiveness with a two-phase (global and local) search [31] while SG [32] combines the local explanation of model [33] along with the symbolic execution [34] to cause an increment both in the discriminatory samples and diversity

## II. BACKGROUND

Adversarial machine learning deals with the study of attacks on machine learning algorithms, and of the defenses against such attacks.[1]. Many years ago, the focus of machine learning engineers and scientist was on obtaining high

accuracy for correct prediction, and while this had greatly been resolve in the past few years. The new challenge had focused on adversarial attack and defense against machine learning model there had been series of research survey which establish the need for protecting machine learning model against various forms of attach which make it to misclassify [2].

During the training of machine learning model, it is usually assumed that the training and test data are generated from the same statistical distribution. This assumption makes the final

model vulnerable to various forms of attack, majority of which includes evasion attacks, [3] data poisoning attacks, [4] Byzantine attacks [5] and model extraction [6].

#### A. Current Adversarial Techniques

1) *Gradient-based evasion attack*: In gradient base evasion attack, a perturbed image which seems like untampered to human eyes is made to be misclassified by neural network model (Fig. 1)[35].

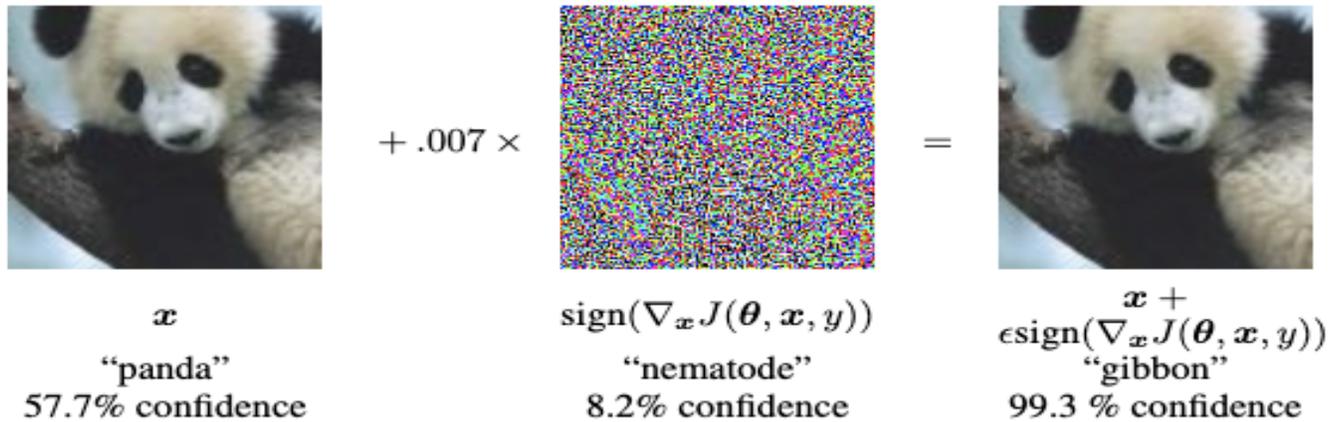


Fig. 1. Adversarial sampling based on addition of perturbation [35]

We can carry out this type of attack by trial and error method as we don't know in advance, the exact data manipulation that will break the model and make it to classify.

Let say we want to probe the boundaries of a machine learning model designed to filter out spam emails, it is possible for us to experiment by sending different emails to see what gets through. And so, a model has been trained for certain words like "momentum", and now we want to make an exceptions for emails that contains other words, if we want to attack, we can craft email with enough extraneous words which will eventually make the model to misclassify it.

#### B. Fast Gradient Sign Method (FGSM)

Let's assume we want to produce an adversarial sample  $x' = x + \eta$  such that  $x'$  is misclassified by the neural network. For us to make  $x'$  and  $x$  produce different outputs,  $\eta$  should be greater than the precision of the features. Let's represent pixel of an image by 8 bits, and we want any information below  $1/255$  of the dynamic range to be discarded. Here, the input is perturbed with the gradient of the loss with respect to the input which gradually increases magnitude of the loss until the input is eventually misclassified.

While  $\epsilon$  decides both the size and sign of each and every element of the perturbation vector which might be matrix or tensor which are being determined by the sign of the input gradient. Here, we just have to linearize the cost function and find the perturbation that maximizes the cost subject to an  $L_\infty$  constraint. This technique causes varieties of models to misclassify input and is also faster than other methods

1) *Projected gradient descent (PGD)*: PGD initializes the sample to a random point in the ball of interest which is being

decided by the  $L_\infty$  norm and does random restarts. This applies the same step as FGSM multiple times with a small step size while at the same time clipping the pixel values of intermediate results of each step to ensure that they are in an  $\epsilon$ -neighborhood of the original image the value of  $\alpha$  used is 1 which means that pixel values are changed only by 1 at each step while the number of iterations were heuristically chosen. This made it sufficient enough for the adversarial example to reach the edge of the  $\epsilon$  max-norm ball.

2) *Carlini and wagner (C&W) attack*: Berkeley, Nicholas Carlini and David Wagner in 2016 propose a faster and more robust method to generate adversarial examples [7]. The attack proposed by Carlini and Wagner begins with trying to solve a difficult non-linear optimization equation. However instead of directly the above equation, Carlini and Wagner propose using a different function and then propose the use of the below function in place of  $f$  using  $z$ , a function that determines class probabilities for given input  $x$ .

With the use of stochastically gradient descent, we can use the above equation to produce a very strong adversarial sample especially when we compare it to fast gradient sign method which can effectively bypass a defensive distillation technique which was previously proposed for adversarial defense [7], [8], [9], [10].

3) *Adversarial patch attack*: Adversarial patch can be devised to fool a machine learning models. They work by causing physical obstruction in an image or by randomizing images with algorithm. Since computer vision models are trained on images that are straight forward. It is inevitable that

any alteration to the input image can make the model to misclassify depending on the severity of the alteration.

We could define a patch function  $p$  corresponding to every transformation  $t \in T$  which applies the transformed patch onto the image and Hadamard product, and the final adversarial perturbed image  $\hat{x}$  which must satisfy  $\hat{x} = p_t(x; \check{z})$  in order to trained patch  $\check{z}$  and some  $t \in T$ .

For us in order to train patch  $\check{z}$ , we could use a variant of the Expectation over Transformations (EOT) framework of Athalye et al. [3]. Let's assume a family of transformations  $T$ , a distance metric  $d$  in the transformed space, and the objective is to find a perturbed image  $\hat{x}$

As the image is expected to be within  $\epsilon$ -ball in anticipation for transformations  $T$ . The attack could find some unconstrained optimization problem. The adversarial patch exploits the way machine learning model are trained for image classification by producing more salient inputs than real world objects. Such salient inputs are misclassified when fed to a machine learning model

### C. Current Defense Strategy and Limitation

1) *Adversarial training*: Adversarial training is a form of brute force supervised learning technique where several adversarial examples are fed into the model and are explicitly labeled as threatening. The approach is similar to a typical antivirus software, which is constantly being updated on a regular basis. As effective as adversarial training may be in defense against adversarial attack, it still requires continuous maintenance or update in order to be effective in combating new threats and it is still suffering from the fundamental problem of the fact that it can only successfully defend against threats or attack that has already happened and is already trained against.

2) *Randomization*: Several adversarial defense methods relied on randomization as a technique for mitigating the effects of adversarial Perturbations in the input and/or feature domain [11]. The idea behind this defense technique is the robustness of deep neural network model to random perturbation. The aim of randomization-based defense is to randomize the adversarial effects of the adversarial sampling into several random effects which is a very ok and normal thing for varieties of deep neural network models.

High success had been achieved by successful defense of Randomization-based defense technique against both black-box and gray-box based attacks, but it is still vulnerable white-box based attack, for example, the EoT method [12] can be easily attacked and compromised simply by considering the randomization process during attack.

3) *Denosing*: In denosing several research works had pointed to both input denosing and feature denosing as a technique for an effective adversarial defense. While input denosing is aimed at complete or partial removal of perturbation from the adversarial samplings or input, feature denosing is aimed at alleviating, reducing or mitigating

effects of adversarial perturbation on important features i.e features that are more impactful on the decision of deep neural network model.

Several methods had been proposed for denosing as a technique for adversarial defense such as conventional input denosing, GAN-based input denosing, auto encoder-based input denosing, feature denosing.

Each of these methods of denosing had been shown to be vulnerable to one form of adversarial attack or another. For instance, Sharma and Chen [13] had shown that input squeezing can bypass by EAD, While good performance was achieved on Testbed by APE-GAN techniques[14], it is easily defeated by white-box based attack[16], As for auto encode-based input denosing, Carlini and Wagner [15],[16] successfully demonstrated that it is vulnerable to the adversarial samples generated by attack, but with feature denosing, research had shown that it merely increase accuracy by 3% which makes it vulnerable to PGD attack.

## III. RESEARCH METHODOLOGY

Fairness when using adversarial sampling as input had shown to cause disparity in accuracy and robustness among different groups [16]. Our method of approach is such as to investigate the cause and offer a solution. The methodologies were in three(s) phases of activities;

### A. Phase-1: Development and Optimization of DNN Model from Scratch

We created a python project in jupyter notebook and created a deep neural network model (DNN) for image classification. We used cifar-10 dataset which consists of 60000 colored images with each image having 32x32 dimensions and categorized into 10 classes of image [Airplane, Automobile, Bird, Cat, Deer, Dog, Frog, Horse, Ship, Truck], with each category containing 6000 colored images. The whole 50000 images in the training dataset folder of cifar-10 was used for training our DNN model while the 10000 images in the test folder of cifar-10 dataset was used to validate our model.

With an initial accuracy of 72%, we needed a higher accuracy and hence, we did some hyper parameter turning by adjusting the learning rate, the number of convolutional layers, and adding some regularization until we are able to get a good accuracy that we can work with, after which we decide to compile and save our new model with KERAS being a high level neural network library that runs on tensorflow.

### B. Phase-2: Generation of Adversarial Sampling

Another python class was created in jupyter notebook in which we wrote algorithms for several adversarial attacks. We wrote adversarial algorithm for adding various kind of noise perturbation to all the images in the training folder (adversarial sampling) (Fig. 2). The algorithm automatically creates new folder and then puts all the adversarial sampling into the new folder. The adversarial folder which contains all the perturbed image or adversarial sampling is named **dogsa** (Fig. 3).

```
def noise_add(path,numb):  
    img = cv2.imread(path,0)  
    im = np.zeros(img.shape, np.uint8)  
    mean=40  
    sigma=50  
    cv2.randn(im,mean,sigma) # create the random distribution  
    noise_image = cv2.add(img, im) # add the noise to the original image  
    io.imwrite("dogsa\\"+str(numb)+".jpg", noise_image)  
    return noise_image
```

Fig. 2. Creation and addition of random gaussian noise distribution to image having image path and position as argument



Fig. 3. Generated adversarial sampling in the dogsa folder after iteration and noise attack

At this point, it was needful for us to test our model with the newly created adversarial sampling to see if it will misclassify those images, having satisfied the criteria of misclassification, it was needful for us to test for fairness across each of the 10 categories of images in the cifar-10 dataset. To actualize this, we separated each category of image in the cifar-10 dataset as separate entity and then observe the accuracy of misclassification across each entity to see if the accuracy of misclassification for each entity will be close, and indeed the accuracy of misclassification for each of the 10 categories of images were closely called which ensure fairness across each group.

### C. Phase-3: Evaluation and Removal of perturbation

This is a very tricky part, as several methods had been proposed with little or no effectiveness. Here, we write algorithm to remove the perturbation, considering the existence of several adversarial attack, we wrote an algorithm to remove all forms of perturbation while at the same trying to maintain the original property of the image. The algorithm iterate through all the images in the adversarial folder where we have our adversarial samplings and remove perturbation in each of them, and in the process creating a new folder name **dogsa-clean** where all the clean images from the adversarial folder are saved (Fig. 4).



Fig. 4. Images in the dogsa-clean folder after passing through the new separation of concern layer from software engineering concept

At this point, we wrote few lines of python code to load our model through keras, and iterating through each of the images in the **dogs-clean** folder where the cleaned images are saved and observe the result. After this, each categories of image were treated as separate entity to account for fairness across each group of images.

#### IV. RESULT AND DISCUSSION

On iterating through each category of adversarial samplings in our adversarial folder to see the rate of misclassification across each image, it was found that there is unfairness as some

of classes have high rate of positive misclassification than the others.

The rate of misclassification was not consistent as Airplane, Automobile, and truck (Fig. 5) has very low rate of misclassification compared with other groups, calini & wagner form of adversarial attack were added to them, while also updating the learning rate and regularization of our initial model and rebuild. The purpose of this is to ensure fairness and consistency across all the classes of image through consistence rate of misclassification (Fig. 6).

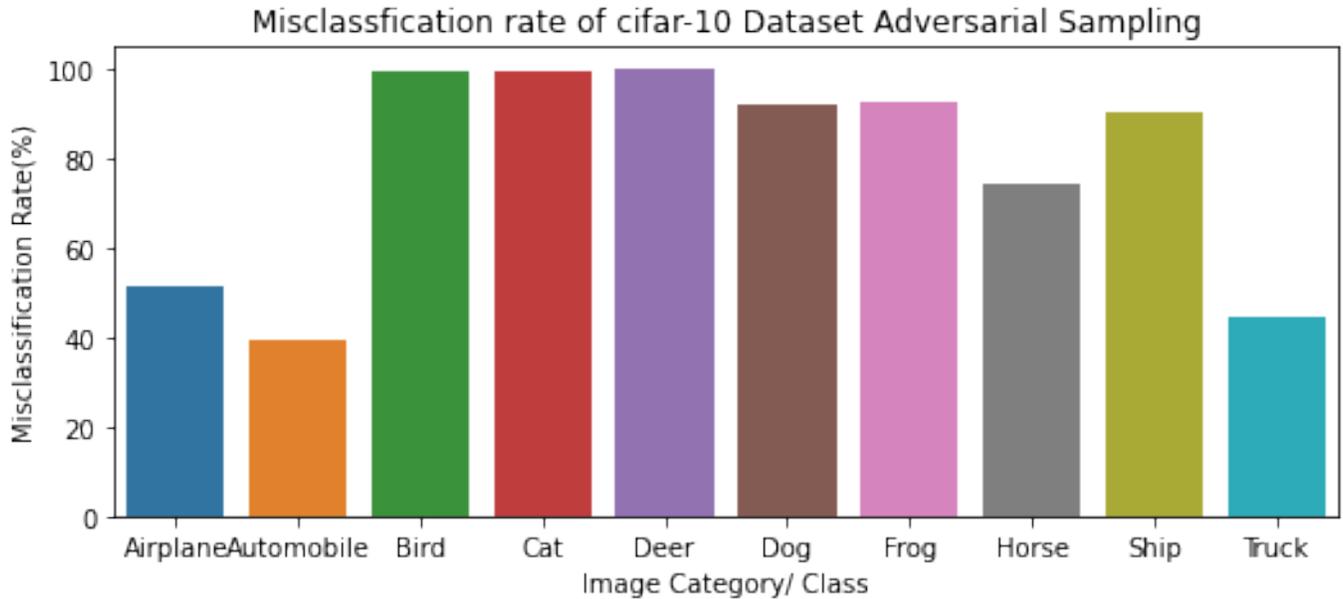


Fig. 5. Plotting of the misclassification rate of the generated adversarial sampling and visualization of fairness across each of the 10 category of images

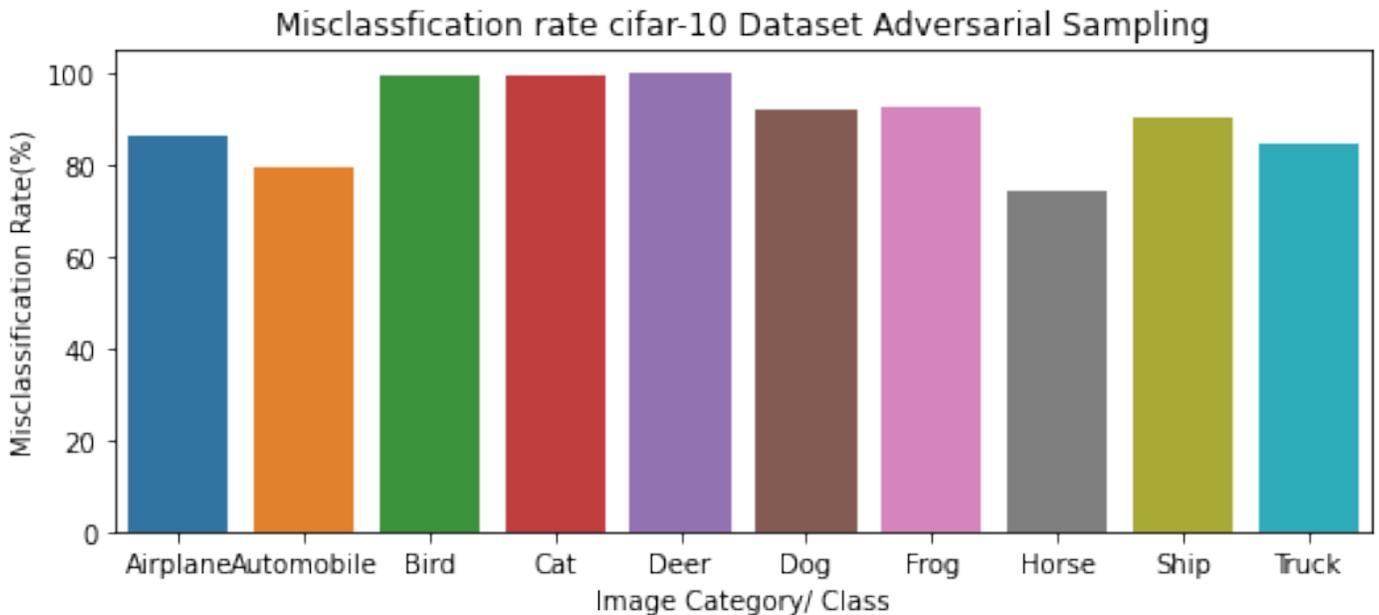


Fig. 6. Plotting of the misclassification rate of the generated adversarial sampling and visualization of fairness across each of the 10 category of images after Calini & wagner attack

With some consistencies in the rate of misclassification, series of algorithms were written to remove greater part of the noises and perturbation for better accuracy. Rather than labeling the attacked images as invalid input, we want the model to be able to predict the attacked images correctly and classify them to the category of images they belonged to.

We wrote an algorithm to iterate through all the images in our adversarial folder, remove as much perturbation and denoise as possible while creating a new folder **dogsa-clean** to store all the new images. Substantial lines of python codes were written to iterate through various classes of cleaned and denoised images in our dogsa-clean folder, and feeding them to the DNN model.

To our surprise and amazement, without making any further hyper-parameter tuning, we were able to have high rate of fairness and consistency across each of the classes of clean images while still maintaining a very good accuracy of prediction in our classifier.

## V. CONCLUSION

In this research, we successfully demonstrate a new method of ensuring fairness across various group of input in deep neural network classifier. Rather than the existing method of training the model on the adversarial sample and label them as invalid. We trained our neural network model on the original image, and without training our model on the perturbed or attacked image. When we feed the adversarial samplings to our model, it was able to predict the original category/ class of the image the adversarial sample belong to.

Through our method we were able to achieve fairness across all the categories of images in the cifar-10 dataset. We also introduce Separation of Concern (SOC) method from full stark software engineering which ensures that we can manage the filter layer as separate entity at any point in the development life cycle without re-training the model.

Surprisingly, we tried to compare the true rate and false rate of fairness for the adversarial sampling across each of the classes of image with that of the cleaned images. We found that the fairness rate was high, consistent and almost the same without any hyper parameter tuning or modification to the filter layer.

## VI. LIMITATION AND FUTURE RESEARCH

In this research, we use the existing forms of adversarial attack for the images. However, we envisage that there will be more sophisticated forms of attack in the future. It is for this reason that we adopt the model of separation of concern from software engineering for our filter layer. In the event of a more robust and sophisticated attack, rather than going back to development to retrain our model, we only need to improve the filter layer.

In addition, the filter layer can be made into a cloud based handy toolbox library. In that case, the filter layer can be managed in the cloud against any future robust attack and be automatically available to all existing deployed model.

## A. Material and Source

We use python 3.9 for this project, pip version 22.3.1, tensorflow, keras to save, load and consume our model and a host of other python libraries.

Our source code for the hypothesis and experiment on this research had been uploaded to github and is made available to the public, and can be accessed through the Uniform Resource Locator (URL) below:  
<https://github.com/IGETOSIN1/Research-Adversarial-Sampling-for-Fairness-Testing>

## REFERENCES

- [1] Kianpour, Mazaher; Wen, Shao-Fang (2020). "Timing Attacks on Machine Learning: State of the Art". Intelligent Systems and Applications. Advances in Intelligent Systems and Computing. Vol. 1037. pp. 111–125. doi:10.1007/978-3-030-29516-5\_10. ISBN 978-3-030-29515-8. S2CID 201705926.
- [2] Jump up to: a b Siva Kumar, Ram Shankar; Nyström, Magnus; Lambert, John; Marshall, Andrew; Goertzel, Mario; Comissoneru, Andi; Swann, Matt; Xia, Sharon (May 2020). "Adversarial Machine Learning-Industry Perspectives". 2020 IEEE Security and Privacy Workshops (SPW): 69–75. doi:10.1109/SPW50608.2020.00028. ISBN 978-1-7281-9346-5. S2CID 229357721.
- [3] Goodfellow, Ian; McDaniel, Patrick; Papernot, Nicolas (25 June 2018). "Making machine learning robust against adversarial inputs". Communications of the ACM. 61 (7): 56–66. doi:10.1145/3134599. ISSN 0001-0782. Retrieved 2018-12-13.[permanent dead link]
- [4] Geiping, Jonas; Fowl, Liam H.; Huang, W. Ronny; Czaja, Wojciech; Taylor, Gavin; Moeller, Michael; Goldstein, Tom (2020-09-28). Witches' Brew: Industrial Scale Data Poisoning via Gradient Matching. International Conference on Learning Representations 2021 (Poster).
- [5] Jump up to: a b c El-Mhamdi, El Mahdi; Farhadkhani, Sadegh; Guerraoui, Rachid; Guirguis, Arsany; Hoang, Lê-Nguyễn; Rouault, Sébastien (2021-12-06). "Collaborative Learning in the Jungle (Decentralized, Byzantine, Heterogeneous, Asynchronous and Nonconvex Learning)". Advances in Neural Information Processing Systems. 34. arXiv:2008.00742.
- [6] Tramèr, Florian; Zhang, Fan; Juels, Ari; Reiter, Michael K.; Ristenpart, Thomas (2016). Stealing Machine Learning Models via Prediction {APIs}. 25th USENIX Security Symposium. pp. 601–618. ISBN 978-1-931971-32-4.
- [7] Carlini, Nicholas; Wagner, David (2017-03-22). "Towards Evaluating the Robustness of Neural Networks". arXiv:1608.04644 [cs.CR].
- [8] "carlini wagner attack". richardjordan.com. Retrieved 2021-10-23.
- [9] Plotz, Mike (2018-11-26). "Paper Summary: Adversarial Examples Are Not Easily Detected: Bypassing Ten Detection Methods". Medium. Retrieved 2021-10-23.
- [10] Wang, Xinran; Xiang, Yu; Gao, Jun; Ding, Jie (2020-09-13). "Information Laundering for Model Privacy". arXiv:2009.06112 [cs.CR].
- [11] Kui Ren, Tianhang Zheng, Zhan Qin, Xue Liu, Adversarial Attacks and Defenses in Deep Learning, Engineering, Volume 6, Issue 3, 2020, Pages 346-360, ISSN 2095-8099, <https://doi.org/10.1016/j.eng.2019.12.012>
- [12] Athalye A, Engstrom L, Ilya A, Kwok K. Synthesizing robust adversarial examples. 2017. arXiv:1707.07397.
- [13] Sharma Y, Chen PY. Bypassing feature squeezing by increasing adversary strength. 2018. arXiv:1803.09868.
- [14] Shen S, Jin G, Gao K, Zhang Y. APE-GAN: adversarial perturbation elimination with GAN. 2017. arXiv: 1707.05474.
- [15] Rokach L. Decision forest: twenty years of research. Inf Fusion. 2016;27:111–25.
- [16] <https://doi.org/10.48550/arxiv.2010.06121>, doi = {10.48550/ARXIV.2010.06121}, url = {<https://arxiv.org/abs/2010.06121>}, author = {Xu, Han and Liu, Xiaorui and Li, Yaxin and Jain, Anil K. and Tang, Jiliang}, keywords = {Machine Learning (cs.LG), Machine Learning (stat.ML), FOS:

- Computer and information sciences, FOS: Computer and information sciences}, title = {To be Robust or to be Fair: Towards Fairness in Adversarial Training}, publisher = {arXiv}, year = {2020}, copyright = {arXiv.org perpetual, non-exclusive license}}
- [17] P. Zhang et al., "Automatic Fairness Testing of Neural Classifiers Through Adversarial Sampling," in IEEE Transactions on Software Engineering, vol. 48, no. 9, pp. 3593-3612, 1 Sept. 2022, doi: 10.1109/TSE.2021.3101478.
- [18] P. Zhang et al., "Automatic Fairness Testing of Neural Classifiers Through Adversarial Sampling," in IEEE Transactions on Software Engineering, vol. 48, no. 9, pp. 3593-3612, 1 Sept. 2022, doi: 10.1109/TSE.2021.3101478.
- [19] F. Schroff, D. Kalenichenko and J. Philbin, "Facenet: A unified embedding for face recognition and clustering", Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp. 815-823, 2015.
- [20] K. Fu, D. Cheng, Y. Tu and L. Zhang, "Credit card fraud detection using convolutional neural networks", Proc. 23rd Int. Conf. Neural Inf., pp. 483-490, 2016.
- [21] E. Wulczyn, N. Thain and L. Dixon, "Ex machina: Personal attacks seen at scale", Proc. 26th Int. Conf. World Wide Web, pp. 1391-1399, 2017.
- [22] Show in Context CrossRef Check for this item at the UTEP Library Google Scholar
- [23] S. Barocas, M. Hardt and A. Narayanan, "Fairness and machine learning", 2019, [online] Available: <http://www.fairmlbook.org>.
- [24] "Draft ethics guidelines for trustworthy AI" in , Brussels, Belgium:European Commission, 2018.
- [25] F. Tramèr et al., "Fairtest: Discovering unwarranted associations in data-driven applications", Proc. IEEE Eur. Symp. Secur. Privacy, pp. 401-416, 2017.
- [26] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger and S. Venkatasubramanian, "Certifying and removing disparate impact", Proc. 21th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, pp. 259-268, 2015.
- [27] O. Bastani, X. Zhang and A. Solar-Lezama, "Probabilistic verification of fairness properties via concentration", Proc. ACM Program. Languages, pp. 118:1-118:27, 2019.
- [28] C. Dwork, M. Hardt, T. Pitassi, O. Reingold and R. S. Zemel, "Fairness through awareness", Proc. Innovations Theor. Comput. Sci., pp. 214-226, 2012.
- [29] S. Garg, V. Perot, N. Limtiaco, A. Taly, E. H. Chi and A. Beutel, "Counterfactual fairness in text classification through robustness", Proc. AAAI/ACM Conf. AI Ethics Soc., pp. 219-226, 2019.
- [30] P. S. Thomas, B. C. da Silva, A. G. Barto, S. Giguere, Y. Brun and E. Brunskill, "Preventing undesirable behavior of intelligent machines", Science, vol. 366, no. 6468, pp. 999-1004, 2019.
- [31] S. Galhotra, Y. Brun and A. Meliou, "Fairness testing: Testing software for discrimination", Proc. 11th Joint Meeting Foundations Softw. Eng., pp. 498-510, 2017.
- [32] S. Udeshi, P. Arora and S. Chattopadhyay, "Automated directed fairness testing", Proc. 33rd ACM/IEEE Int. Conf. Automated Softw. Eng., pp. 98-108, 2018.
- [33] A. Aggarwal, P. Lohia, S. Nagar, K. Dey and D. Saha, "Black box fairness testing of machine learning models", Proc. ACM Joint Meeting Eur. Softw. Eng. Conf. Symp. Foundations Softw. Eng., pp. 625-635, 2019.
- [34] M. T. Ribeiro, S. Singh and C. Guestrin, "", Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, pp. 1135-1144, 2016.
- [35] Goodfellow, Ian J., Jonathon Shlens and Christian Szegedy. "Explaining and Harnessing Adversarial Examples." *CoRR* abs/1412.6572 (2014): n. page

# Text2Simulate: A Scientific Knowledge Visualization Technique for Generating Visual Simulations from Textual Knowledge

Ifeoluwatayo A. Ige<sup>1</sup>, Bolanle F. Oladejo<sup>2</sup>

Computer Science Department, Rochester Institute of Technology, Rochester, USA<sup>1</sup>  
Computer Science Department, University of Ibadan, Ibadan, Nigeria<sup>1,2</sup>

**Abstract**—Recent research has developed knowledge visualization techniques for generating interactive visualizations from textual knowledge. However, when applied, these techniques do not generate precise semantic visual representations, which is imperative for domains that require an accurate visual representation of spatial attributes and relationships between objects of discourse in explicit knowledge. Therefore, this work presents a Text-to-Simulation Knowledge Visualization (TSKV) technique for generating visual simulations from domain knowledge by developing a rule-based classifier to improve natural language processing, and a Spatial Ordering (SO) algorithm to solve the identified challenge. A system architecture was developed to structurally model the components of the TSKV technique and implemented using a Knowledge Visualization application called ‘Text2Simulate’. A quantitative evaluation of the application was carried out to test for accuracy using modified existing information visualization evaluation criteria. Object Inclusion (OI), Object-Attribute Visibility (OAV), Relative Positioning (RP), and Exact Visual Representation (EVR) criteria were modified to include Object’s Motion (OM) metric for quantitative evaluation of generated visual simulations. Evaluation for accuracy on generated simulation results were 90.1, 84.0, 90.1, 90.0, and 96.0% for OI, OAV, OM, RP, and EVR criteria respectively. User evaluation was conducted to measure system effectiveness and user satisfaction which showed that all the participants were satisfied well above average. These generated results showed an improved semantic quality of visualized knowledge due to the improved classification of spatial attributes and relationships from textual knowledge. This technique could be adopted during the development of electronic learning applications for improved understanding and desirable actions.

**Keywords**—Knowledge visualization; visual simulation; text-to-simulation knowledge visualization technique; natural language processing; electronic learning

## I. INTRODUCTION

Knowledge visualization is the application of visualization techniques to disseminate knowledge among individuals [1], [2]. The main purpose of knowledge visualization is to improve the communication of knowledge through visual means. Although most available knowledge sources are in either numeric or textual formats, it is highly necessary to provide a visual representation of such knowledge for easier assimilation and retention in human minds and for fostering required action [3]. Visual knowledge could be represented

with charts, maps, images, scenes, simulations, and videos. In this article, a textual knowledge visualization technique that visually simulates textual knowledge was developed. A visual simulation is a visual representation containing visual objects that dynamically move based on predefined spatial attributes or collisions. Although visual metaphors, scenes, and animations could be used for the visualization of textual knowledge, these techniques do not emphasize precision in the visual representation of spatial attributes. Such techniques may not be appropriate for related domains that require accuracy in visual object positioning and movement. Furthermore, expressing textual knowledge in visualizations is of utmost importance for effective assimilation to be achieved [4]. This paper, therefore, presents a Text-to-Simulation Knowledge Visualization (TSKV) technique for generating visual simulations from domain knowledge. A review of related works is described in the next section of this article which is followed by an implementation of the TSKV technique. A quantitative evaluation is finally presented to test for accuracy using modified existing information visualization evaluation criteria.

## II. REVIEW OF RELATED WORK

Knowledge visualization, according to [5] and [6] can be summarized into four major formats. Sketches, diagrams, images, and interactive visualizations. Fig. 1 show a word cloud of knowledge visualization techniques. A sketch is a visual representation that shows an abstract drawing or prototype of an idea or concept. Diagrams consist of boxes and circles that represent concepts and entities; and lines, and edges that depict the relationship between the entities. Diagrams are used to illustrate the classification and clustering of related concepts in a domain. Examples are Knowledge Graphs used for visualizing search histories [7]; concept maps used to illustrate concepts and relationships between them in specific domains such as medicine [8]; teaching and learning among others [9]. An image (visual metaphor) is a pictorial representation of humans and events. It is usually generated through rendering, photography, or paintings. Images are used to express emotions and give an idea about a concept such as comic images, icons, and emojis for storytelling [10]. Interactive visualization involves visually representing knowledge with animated objects and shapes. It allows users to interact and make decisions while viewing the visualization. Visualizations are usually shown in an ordered sequence of images. Examples are models such as archetypes [11];

semantic visualization tools [12]; virtual reality [13]; simulations, and animation [14].



Fig. 1. Knowledge visualization techniques

Existing literature research has shown that explicit knowledge could be visually represented with knowledge graphs, knowledge maps, and concept maps. Hao [15] used knowledge graphs to develop a surveying and remote-sensing application. Li [16] employed a concept map to develop adaptive learning systems. Visual metaphors replace key terms and concepts found in textual documents which visual characters. Huron [17] used data feeds to represent sediments. Chau [18] also used flowers to represent search results from the web. Hiniker [19] applied visual metaphors for clustering and viewing citations of literature from a large database. Keith [20] used visual metaphors to develop narrative maps for visualizing online narratives.

More recent KV techniques such as text-to-scenes, text-to-videos, and text-to-simulations generate more realistic visual presentations in form of images [21]; animations [22]; scenes [23], [24]; videos [25] and visual simulations [14]. A brief discussion of the major techniques related to our work is presented thus:

#### A. Text-to-Scene

Text-to-scene conversion generates static scenes from natural language text. Several authors have worked on this research area by applying different AI approaches. The first conversion system was WordEye [26]. Other authors applied machine learning techniques for text-to-scene conversion [27], [28]. Deep learning techniques such as Variational Auto Encoders (VAE) and Generative Adversarial Networks (GAN) are recently being applied for text-to-scene conversion [29], [30]. Our technique does not produce static scenes but dynamic simulations showing object-to-object collision behavior and movement.

#### B. Text-to-Simulation

Visual simulations are generated from textual inputs. Very limited research in this area exists in [14], [31], [32], and [33]. VoxSim architecture was presented in their work. They

applied a rule-based approach specifically predicate logic to set rules for each motion verb which was applied during conversion. However, precise geometric information of objects and their relative positions with other objects were not specified. Also, behavioral attributes of objects such as motion were not specified. The above issues were reported to have led to ambiguities in some simulation results. In our work, precise spatial grounding information of visual objects was included. Exact geometric information and their behavior attributes for all visual objects was also defined.

#### C. Text-to-Video

This involves converting natural language text to Videos that semantically depict the textual input. Recent research on text-to-video conversion exist [25], [34], [35]. A.I. techniques have been applied. Specifically, machine learning [36]; deep learning and neural networks such as convolutional neural networks [34]; recurrent neural networks [37], long short-term memory networks [38], and generative adversarial networks [25], [35] Although a rule-based approach was adopted, there is a slight resemblance between our work and most text-to-video conversion systems in that our prototype tool coupled with TTS and an existing screen capturing tool can generate a video of all simulations performed.

#### D. Natural Language Processing (NLP)

This involves NLP which involves several tasks such as sentence segmentation, tokenization, Part-of-Speech (POS), Named Entity Recognition Dependency parsing among others [39]. SpaCy [40], Natural Language Processing Tool Kit [41] and Stanford CoreNLP tool [42] are some NLP tools that perform NLP tasks. Although these tools perform these tasks, classification and extraction of domain-related keywords and attributes remain a research challenge due to the semantic interpretation of such words. A rule-based classifier was developed to classify and extract domain-related keywords, keyword attributes, and relations between keywords.

#### E. Spatial Relations and Arrangement

This involves defining the specific mapping of spatial keywords to predefined spatial information. Chang [28] and Ma [24] developed spatial relations for prepositions. Fisher [43] presented an arrangement model for determining the order of object placement and position. Our work also defines specific relation information for spatial keywords and prepositions by developing an algorithm for spatial positioning and rendering of objects.

#### F. CAD Models and Scenes

Computer-Aided Design is employed for creating models which are visual replicas of real-world objects. Most KV applications make use of existing CAD model datasets such as ShapeNet [23]; ScanNet [44] and Scene datasets such as SceneNet [45] for developing visual representations of textual knowledge. This work did not make use of existing datasets but developed new models. This is due to the scarce availability of models in the selected domain used for application validation.

### III. METHODOLOGY

We present a mathematical model for generating visual simulations from textual knowledge and a system architecture in the subsequent subsections.

#### A. Mathematical Model Formulation

We mathematically model the task of generating visual simulations from textual knowledge. This is carried out using linear functions given a set of textual inputs  $I_f (i = 1, 2, \dots, h, n)$ . The following definitions show each milestone required from textual input to visual simulation output.

**Definition 1:** Suppose there exist a Semantic Classifier  $C$ ;

$$C(I_f) = [\{E\}, \{A\}, \{e_f R e_h\}] \quad (1)$$

Where the set of entities is  $E_f (i = 1, 2, \dots, h, n)$ ,  $A$  is the attribute(s) set;  $R$  is the relationship between entities,  $e_f$  and  $e_h$  are entities.

**Definition 2:** Suppose there exist a model repository  $O$  and entity set  $E$ ;

$$o_f = f(e_f) \quad (2)$$

Where  $o$  and  $e$  are some models and entities in  $O$  and  $E$  and all entities in  $E$  have a one-to-one function in  $O$

**Definition 3:** Suppose there are a finite set of shapes  $O_f (i = 1, 2, \dots, h, n)$  with scaling factor  $s_f$ ; centroid coordinate  $p(b_f, c_f)$ , spatial attributes  $A_f$  and relation  $r$ . Let  $m$  be some numeric value;  $\alpha$  be the length of  $o_f$  and  $\alpha > A_i$ . Let there exist Scaling and relative positioning Transforms  $S$  and  $R$  respectively. Let there exist a rendering Engine  $K$ . The visual simulation is given by  $V$ .

1) For single object placement  $o_f$ ;

$$f(o_f) = S + p(b_f, c_f) \quad (3)$$

$$V = K(f(o_f)) \quad (4)$$

2) For two relatively positioned objects  $o_f$  and  $o_h$ ;

$$f(o_f) = S_i + p(b_f, c_f)$$

$$f(o_h) = S_h + p(b_h, c_h)$$

$$V = K(f(o_f \cap o_h)) \quad (6)$$

Where:

$$p(b_f, c_f) = p[(b_h), (m + c_h)] \quad \text{and}$$

$$f(o_f \cap o_h) = S_f + p(b_f, c_f) + S_h + p(b_h, c_h) + R(r)$$

3) For two relatively positioned objects  $o_f$  and  $o_h$  with spatial attribute  $A_i$ ;

$$f(o_f) = S_f + p(b_f, c_f); f(o_h) = S_h + p(b_h, c_h)$$

$$V = K(f(o_f \cap o_h)) \quad (7)$$

Where:

$$p(b_f, c_f) = p[(b_h - (\alpha/2 - A)), (m + c_h)] \quad \text{and}$$

$$f(o_f \cap o_h) = S_f + p(b_f, c_f) + S_h + p(b_h, c_h) + R(r)$$

Given textual input, the semantic classifier extracts all entities, attributes, and relationships between entities from textual input as described in Eq. (1). Next, a set of entities are extracted from the model repository (Eq. (2)). Each model has a unique centroid coordinate  $p(b_h, c_h)$  and a scaling vector. Eq. (3) and (4) describe how single models are rendered while taking into consideration its centroid coordinate and its scaling vector. To render two objects relative to each other (Eq. (6)) the centroid coordinate of an object is recalculated using the centroid coordinate of its relative object  $b_h$  and some numeric value which could be the height or thickness of the relative object added to  $c_h$  to get the  $c_h$ . For object relative positioned to each other with spatial attributes (Equation 7), the  $b_f$  value of an object is recalculated by; first subtracting the value of the spatial attribute from the mid-length value of its relative object. This result is further subtracted from the  $b_h$  value of its relative object. Some numeric value which could be the height or thickness of the relative object added to  $c_h$  to get the  $c_f$ .

#### B. System Architecture

Fig. 2 shows the Text-to-Simulation Knowledge Visualization architecture. It comprises four major modules: Natural Language Processing module, 2D Graphic Models Knowledge Base, Spatial Ordering module, and the user interface module.

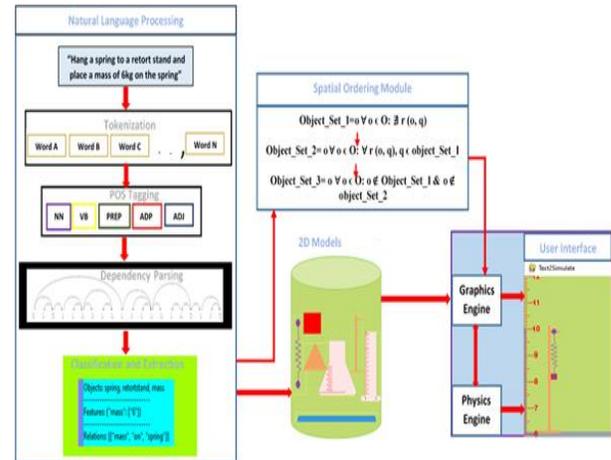


Fig. 2. Text-to-simulation knowledge visualization (TSKV) architecture

The Natural Language Processing module performs natural language processing tasks which are tokenization (breaking down the user's input sentence into words, numbers, punctuation marks, full stops, and discrete items); Part of Speech Tagging (allocating a POS tag per word); Dependency Parsing (assigning dependency labels to show relationship patterns between object and subject tokens); Classification and Extraction of domain-specific words, attributes, and relationships. 2D models can be found in the image repository. The Spatial Ordering Module determines the order of model rendering. The Graphic and Physics Engines are built into the application developed to cater for collision and interaction among models and also for rendering.

1) Natural language processing and dependency parsing:

We make use of SpaCy toolkit for natural language processing and dependency parsing. It is an existing industrial natural language processing library written in Python programming language [40]. SpaCy is known for a higher level of speed and accuracy in major NLP tasks such as POS tagging [46] and dependency parsing [47] when compared to other NLP applications. Tokenization of text, POS tagging, and parsing sub-libraries are used in our work. The NLP pipeline selected for this research supports English Language since it is the commonest medium of communication where this architecture was implemented and evaluated.

Each word in the user’s input is passed through a text corpus for relevance. If the sentence is not domain-related, the user is prompted to input domain-related sentences. Tokenization is done by converting each word in the sentence to tokens. Dependency parsing is done by assigning POS tags to each word, assigning dependency labels that show relationship patterns between object and subject tokens. Fig. 3 shows an example of a dependency graph given the sentence: *place a ruler*”

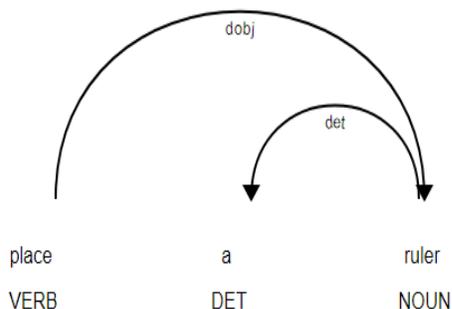


Fig. 3. Dependency graph for ‘place a ruler’

2) 2D models: A repository was created to store 2D models. Major apparatus objects used for performing High school experiments in Mechanics (a subtopic in Physics) are modeled in 2D for this research. The intended and purposely selected users for evaluation of the implemented architecture informed the choice of 2-Dimensional modelling. 2D graphic models provide necessary visual knowledge without much distraction and complexity.

The apparatus image library consists of 14 classes as shown in Fig. 4.

Each 2D apparatus model was created using modified existing objects from the Pymunk library. An apparatus model comprises of one or more objects. Fig. 4 is a class diagram showing each apparatus class, the attributes, and the methods. It also shows the relationship among the apparatus classes. The classes are CircularWeight(), BlockWeight(), RetortStand(), SpiralSpring(), KnifeEdge() and MetreRule(), Table(), Cup(), InclinedPlane(), TestTube(), Beaker(), GraduatedCylinder() and ErlenmeyerFlask().

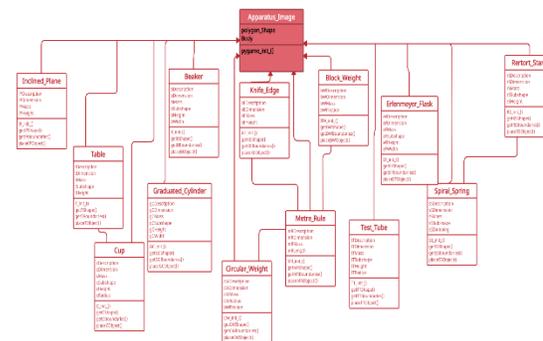


Fig. 4. Class diagram for apparatus model repository

Thirteen sub-classes inherit attributes ‘polygon\_shape’ and ‘body’ from the main class ‘Apparatus\_Image’ as shown in Fig. 4. Each object class contains the following methods:

- a) *\_init\_()*: This is a constructor which is called when a model is instantiated. It is used to initialize each class’s attributes.
- b) *getShape ()*: This method returns the model when called
- c) *getBoundaries ()*: This returns the dimensions of each model (left, right, top, bottom, centroidCordinate(x,y))
- d) *placeObject ()*: This places the model in space when called.

3) *Rule-based classification and extraction algorithm*: We present a novel rule-based classification algorithm that accepts a predefined objects-of-interest list and tokens (Fig. 5). Extracted noun tokens from the POS tagging must exist in the predefined list of objects of interest. The algorithm returns an object list, objects-attribute list, relations list, and object-rel-object list which shows the relationship between objects. In this work, the relationship between two objects is only considered.



Fig. 5. Rule-based classification and extraction algorithm

4) *Attributes and relations grounding*: We consider spatial attributes such as the weight, size, and unique position of objects which are extracted during NLP. The object's color and other characteristics are not emphasized as much work has been done on this area in literature [24], [28]. The spatial attributes are very essential for precision in visual simulation output. Three predefined sets of relations: *on*, *under*, and *inside* are considered. The bounding box approach was employed to determine close distance and positions appropriately. The (x, y) axis coordinates, the height, and the width of the bounding box of the model were used to determine relative positions.

5) *Spatial Ordering*: Spatial ordering is the task of sequentially rendering models and determining relative positioning for related models. Given a set of object list O, object-attribute list P, relations list Q, and object-rel-object list R, a visual simulation is sequentially rendered based on the following procedure:

$$\forall o \in O, p \in P, q \in Q, r \in R;$$

Render Object\_Set\_1 (s);

Render Object\_Set\_2 (s)

Render Object\_Set\_3 (s)

Where:

a) Object\_Set\_1\_list = o  $\forall o \in O: \nexists r(o, q) // o$  is independent of q

b) Object\_Set\_2\_List = o  $\forall o \in O: \forall r(o, q), q \in$  Object\_Set\_1

c) Object\_Set\_3\_List = o  $\forall o \in O: o \notin$  Object\_Set\_1 & o  $\notin$  Object\_Set\_2.

The next section presents an implementation of the methodology.

#### IV. "TEXT2SIMULATE" KNOWLEDGE VISUALIZATION SYSTEM

We develop a knowledge visualization system called *Text2Simulate* based on the architecture shown in Fig. 2.

##### A. Software Requirement Gathering and Elicitation

Domain experts (three teachers) who taught physics in High schools located in a remote area were interviewed to retrieve system requirements since the targeted users of the application are high school physics subject tutors and senior high school students from remote areas. It is expected that the users have an elementary level of proficiency in English language. The users should also be familiar with basic concepts and terms used in high school physics. The *Text2Simulate* Knowledge visualization system can be used by students with little or no supervision of teachers.

##### B. Dataset

The dataset used for this application is the models in the apparatus image library and physics-related sentences in natural language text. The apparatus object image library contains apparatus object images. Existing objects in the Pymunk library are modified and used to create apparatuses models for Ruler, knife-edge and other apparatuses shown in Fig. 2.

##### C. Graphic User Interface

The user interface is divided into three major sections as shown in Fig. 6. The visual simulation is viewed in the right section. On the upper left, the textbox is used for accepting textual input from users. Apparatus models can also be viewed by clicking on objects from the Toolbar. Selected models are then viewed on the right viewing pane. When the button 'analyze' is clicked, the extracted objects, object-attribute list, and the object-rel-object list are shown on the lower-left pane. The user then clicks the 'simulate' button to generate visual simulations which are shown on the right pane.

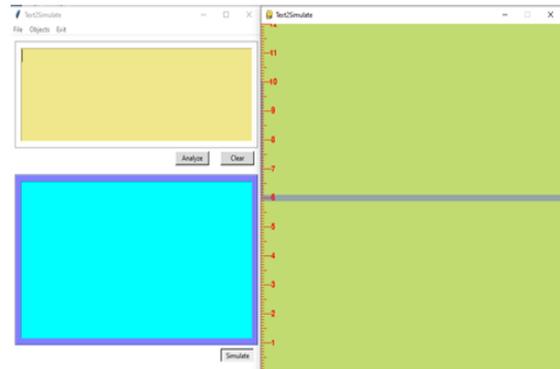


Fig. 6. Text2Simulate knowledge visualization system interface

##### D. Implementation Results

The models can be viewed by clicking each model as seen in Fig. 7. A table model is shown when clicked. Fig. 8 to 10 are screenshots of sentence inputs, classification results and visual simulation. The classification results show the objects, attributes and relations extracted from the sentence input. The figures show the screenshot of the visual simulation of the sentence. Fig. 8 shows a visual simulation result for 'Hang a spring to a retort stand and place a mass of 6kg on the spring'. During visual simulation, the motion can be viewed when the mass is attached to the spring. The spring continues to oscillate till it gets the equilibrium position. The screenshot was taken when the spring got to its equilibrium position.

Fig. 9 and 10 show visual simulation for sentence inputs describing principle of moments using a balanced ruler experiment [48].

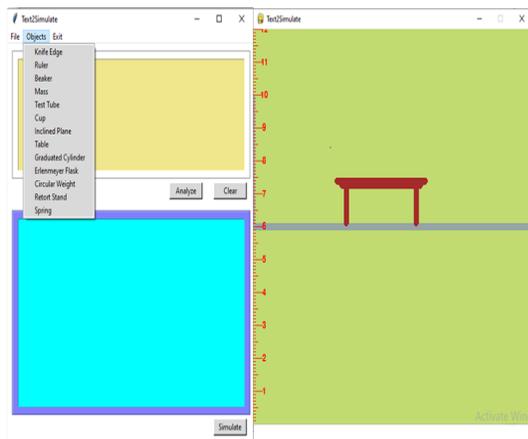


Fig. 7. A visual representation for 'Table' model

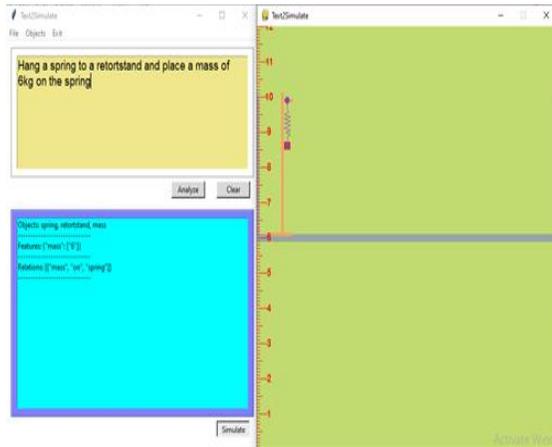


Fig. 8. Visual simulation screenshot of 'Hang a spring to a retort stand and place a mass of 6kg on the spring' sentence when spring is at equilibrium position

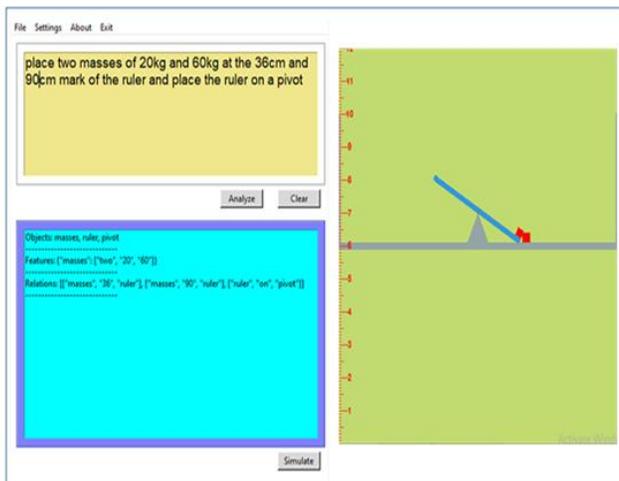


Fig. 9. Visual simulation screenshot of 'place two masses of 20kg and 60kg at the 36cm and 90cm mark of the ruler and place the ruler on a pivot'

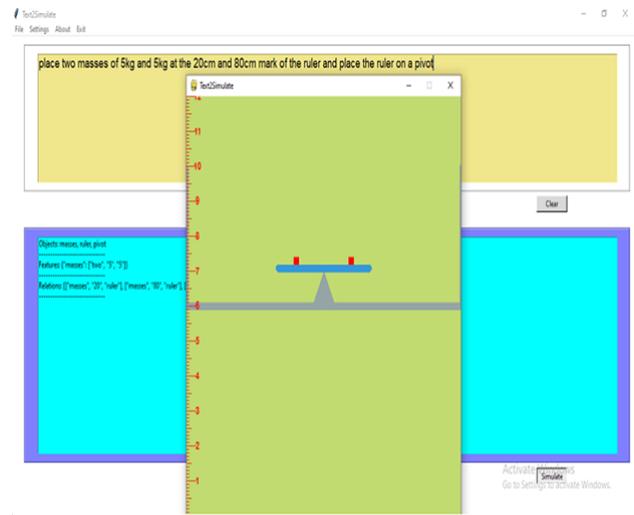


Fig. 10. Visual simulation screenshot of 'place two masses of 5kg and 5kg at the 20cm and 80cm mark of the ruler and place the ruler on a pivot'

## V. QUANTITATIVE EVALUATION

We first perform a quantitative evaluation on the rule-based semantic classifier. Then, the accuracy of converting textual knowledge to visual simulation representation is evaluated. Finally, a user evaluation on the knowledge visualization tool is conducted.

### A. Rule-Based Classification Evaluation

A total of 110 sentences were purposively selected and used for evaluation. A total of 60 sentences were extracted from the general domain; while the remaining 50 were domain-specific sentences. *Objects\_of\_Interest*, *Object\_Attribute* and *Object\_Relation\_Object* classification results are compared with human-generated classification. Standard performance evaluation metrics (recall, accuracy, precision, and F1 score) based on the confusion matrix are employed. Each outcome is assigned True Positive (TP) if it correlates with human-generated classification; assigned False Positive (FP) if it is extracted as a member of a list but false with human-generated classification and assigned False Negative (FN) if human-generated classification assigns it to a list but is not included in the extracted list. Eftimov [49] and Popovski [50] reported that the True Negative metric is not required for the evaluation of rule-based entity classification methods. Hence, True Negative values are not reported. Table I presents the three classification categories; Objects of Interest list; *Object-Attribute* list and *Object-Relation-object* list.

TABLE I. THE CONFUSION MATRIX FOR EVALUATION OF OBJECTS, ATTRIBUTE AND RELATIONS CLASSIFICATION

	LIST	TP	FP	FN
1	Objects of Interest	242	0	11
2	Object-Attribute	32	0	4
3	Object-Relation-Object	52	0	4

Fig. 11 displays the graph showing recall, accuracy, and F1 score values for *Objects\_of\_Interest*, *Object\_Attribute*, and *Object\_Relation\_Object* classification.

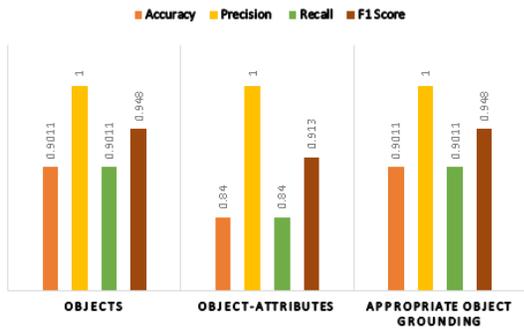


Fig. 11. A graph showing performance evaluation metrics for *Objects\_of\_Interest*, *Object\_Attribute* and *Object\_Relation\_Object* classification

The accuracy for *Objects\_of\_Interest* classification is 0.9565 as presented in Fig. 11; precision is 1; recall is 0.9565 and F1 score of 0.9777. For *object\_Attribute* classification and extraction, the accuracy is 0.8888; precision is 1; recall is 0.8888 and F1 score is 0.9411. Evaluation for the *Object\_Relation\_Object* classification produces; 0.9286 (recall); 1 (precision); 0.9286 (accuracy); and F1 score of 0.9630. The precision value of 1 for all the classifications is achieved because objects belonging to the list from the classifier actually belong according to human judgment. It can also be attributed to the advantage of employing a rule-based approach for the classifier modeling as reported in Al-Moslmi [51]. Accuracy and recall have the same value for all the categories. This shows that the classifier can correctly classify and extract all three categories to separate lists. This is also because True Negative outcome is not computed for this evaluation. Summarily, it can be concluded that evaluation results show that the classifier performed well above average. During error analysis of semantic classifier results, the dependency parser did not identify a few relations existing between objects as a preposition. They were identified as modifiers. However, human judgment categorized them as prepositions. Some attributes were also not identified since they did not have numeric values. The dependency parser did not identify a few compound words as nouns (objects). The words were split into modifier and noun ('inclined plane). However, human classification categorized them as Nouns.

### B. Evaluation of Knowledge Visualization System based on Visual Simulation Results

Existing evaluation criteria used for evaluating the visual representation of information visualization techniques in [52] and [53] were adopted and modified to include Object's Motion criteria. Two domain experts were selected to evaluate *Text2Simulate* Knowledge Visualization application based on visual representation results. The simulation generated from the *Text2Simulate* application were evaluated using 50 purposively selected domain-specific sentences and compared with human judgment based on the following modified criteria: Objects Inclusion (OI); Object-Attributes Visibility (OAV); Object's Motion (OM); Relative Positioning (RP) and

Exact Visual Representation of Text (EVR). Each criterion was evaluated using performance metrics based on a confusion matrix. Each result is Assigned True Positive (TP) if the visual simulation matches with human judgment; Assigned False Positive (FP) if the visual simulation is not semantically correct based on human judgment and Assigned False Negative (FN) if there is no visual simulation (object is static) but there should be simulation-based on human judgment. The confusion matrix for visual simulation tasks is presented in Table II.

TABLE II. CONFUSION MATRIX OUTCOMES FOR VISUAL SIMULATION

	CRITERIA	TP	FP	FN
1	Objects Inclusion (OI)	82	0	9
2	Object-Attributes Visibility (OAV)	21	0	4
3	Object's Motion (OM)	82	0	9
4	Relative Positioning (RP)	27	0	3
5	Exact Visual Representation of Text (EVR)	48	1	1

Fig. 12 shows the evaluation results for precision, accuracy, recall, and F1 score metrics of Objects Inclusion, Object-Attributes, Object's Motion, and Relative Positioning.

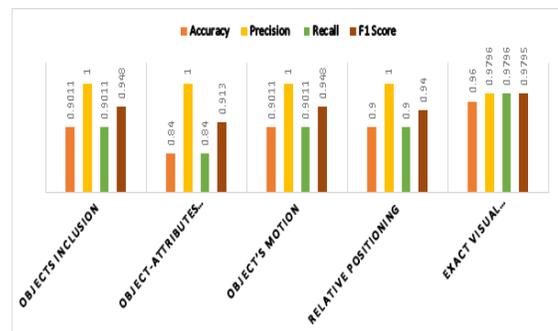


Fig. 12. A graph showing performance evaluation metrics for visual simulation

The visual presentation results showed almost a hundred percentage positivity for the knowledge presented as shown in Fig. 12. Spatial attributes were significantly visualized as well as the motion of objects which was dependent on their unique attributes, sizes, and intersection with other relative objects. During error analysis, a statement such as 'place two masses of 5kg and 20kg on a table' produced a visual representation showing only one mass and a table. Although the second mass was also rendered, it was hidden as both masses were placed at the same default location. Also, it was observed that some objects were not rendered even though the classifier identified them as objects as their models were not found in the image library.

### C. User Study Evaluation

A total of 10 physics teacher participants and 20 student participants who have little or no prior graphic knowledge and reside in developing areas were purposively selected for the user study. Both set of participants were trained on how to simulate physics experiments on the knowledge visualization tool. They were then asked to perform two experiments

namely: Principle of Moments Using a Balanced Ruler and Hooke's Law as shown in Fig. 8, 9 and 10. The participants were finally asked to fill a user feedback questionnaire which was based on a two-point Likert scale (Yes/No) after the performing both experiments. The questions in the survey was drafted to indicate the level of system effectiveness and their overall satisfaction with the application developed based on the following 10 purposively selected metrics: Graphic design (T1), User-Friendliness (T2), Meaningful Arrangement (T3), Meaningful Size (T4), Object-Attribute visible (T5), Semantic Correlation of Text and Simulation (T6), Ability to Understand (T7), Ease of Use (T8), Reading Robustness (T9), Reusability (T10). *Text2Simulate* knowledge Visualization application was chosen as the independent variable while T1 to T10 were the dependent variables. Results of the analysis of mean is in Fig. 13.



Fig. 13. Graph of mean for teachers and students evaluation of *Text2Simulate* based on T1-T10 metrics

Fig. 13 showed that the teachers were fully satisfied as T2, T3, T4, T5, T6, T8, T9, T10, and T11 metrics have an average value of 1. The metrics are User-Friendliness, Meaningful Arrangement, Meaningful Size, Object-Attribute visible, Semantic Correlation of Text and Simulation, Ease of Use, Reading Robustness, Reusability, and Knowledge Sharing. This can be attributed to the essential need for *Text2Simulate* application as a teaching aid and for electronic learning due to the recent pandemic occurrence as stated by one of the teachers. The high performance on these metrics is also due to the current non-availability of the laboratory apparatuses in schools of the participants and their preference for *Text2Simulate* application over conducting the experiment for students in their laboratories if available. The mean scores for student's overall satisfaction based on the metrics are well above 0.5. It is shown that the scores range from 0.7 to 0.95. This reflects that most of the students were satisfied with the application.

## VI. CONCLUSIONS AND FUTURE WORK

A Text-to-Simulation Knowledge Visualization (TSKV) technique for generating visual simulations from domain knowledge has been presented and implemented using a newly developed Knowledge Visualization application called '*Text2Simulate*'. The generated results have shown that precise semantic visual representations of spatial attributes and relationships between objects of discourse can be generated from natural language text using the above

technique. The developed rule-based semantic classifier can be used for domain-related classification of text which requires classification and extraction of objects, object properties, and the relationship between objects. The text-to-simulation technique for knowledge visualization produced a better visual representation of textual knowledge than existing knowledge visualization techniques due to its ability to visualize spatial object attributes retrieved from the text. This technique could be employed when developing electronic learning applications.

## REFERENCES

- [1] R.A. Burkhard. Towards a Framework and a Model for Knowledge Visualization: Synergies between Information and Knowledge Visualization. *Knowledge and Information Visualization. S.-O. Tergan and T. Keller. Berlin, Springer Verlag, 2005: 226-243.*
- [2] X. Bai, L. Li and S. Zhang. Software for 3D model retrieval using local shape distributions. <http://code.google.com/p/shape-retrieval>. 2012.
- [3] J. Sleight, M. Schneider, J. Amann and E. Vayena. Visualizing an Ethics Framework: A Method to Create Interactive Knowledge Visualizations from Health Policy Documents. *J Med Internet Res* 2020: 22(1).
- [4] M.J.Eppler. What is an Effective Knowledge Visualization? Insights from a Review of Seminal Concepts. *2011 15th International Conference on Information Visualization, 349-354.*
- [5] M. Eppler and B. Burkhard. Knowledge Visualization. In *Schwartz, David G. (ed.): Encyclopedia of Knowledge Management*. Hershey, PA: Idea Group Reference, 2005: 551-560.
- [6] R. Meyer. Knowledge Visualization. *Trends in Information Visualization, 2010: 23.*
- [7] L. Xu, Z. T. Fernando, X. Zhou and W. Nejd. LogCanvas: Visualizing Search History Using Knowledge Graphs. In *Proceedings of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. 2018.
- [8] R. A. Burkhard. Learning from Architects: Complementary Concept Mapping Approaches. *Information Visualization, 2006; 5(3):225-234. doi:10.1057/palgrave.ivs.9500128*
- [9] K. Shatri and K. A. Buza. The Use of Visualization in Teaching and Learning Process for Developing Critical Thinking of Students. *Eur J Soc Sci Edu Res* 2017; 9: 71-74.
- [10] C. Tong, R. C. Roberts, R. Borgo, S. Walton, R. Laramée, K. Wegba, A. Lu, Y. Wang, H. Qu, Q. Luo, and X. Ma. Storytelling and Visualization: An Extended Survey. *Information, 2018; 9 (65).*
- [11] H.A.M. Sasieta, F. D. Beppler and R. C. Pacheco. A Model for Knowledge Visualization Based on Visual Archetypes. *Acta Scientiarum Tech, 2012; 34(4): 381-389.*
- [12] G. Molcho and R. Schneor. MIND - Semantic based Knowledge Visualization. In *Procedia CIRP* 2015; 36: 89-94.
- [13] Z. Qu, S. Hou, L. Zhu, J. Yan and S. Xu. The Study of Smart Grid Knowledge Visualization Key Technologies. *Telk Indo J Elec Eng* 2014; 12 (1): 323-333.
- [14] J. Pustejovsky and N. Krishnaswamy. Building Multimodal Simulations for Natural Language. *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Tutorial Abstracts*. Association for Computational Linguistics, Spain, 2017.
- [15] X. Hao, Z. Ji, X. Li, L. Yin, L. Liu, M. Sun, Q. Liu and R. Yang. Construction and Application of a Knowledge Graph. *Remote Sens* 2021; 13: 2511. <https://doi.org/10.3390/rs13132511>
- [16] Y. Li, Z. Shao, X. Wang, X. Zhao, and Y. A. Guo. Concept Map-Based Learning Paths Automatic Generation Algorithm for Adaptive Learning Systems. *IEEE Access* 2019; 7: 245-255.
- [17] S. Huron, R. Vuillemot and J. D. Fekete. Visual sedimentation. *IEEE Trans. Vis Comp Graphics* 2013; 19: 2446-2455
- [18] M. Chau. Visualizing web search results using glyphs: Design and evaluation of a flower metaphor. *ACM Trans Mgt Info Sys* 2011; 2(1):2.

- [19] A. Hiniker, S. Hong, K. Kim, N. Chen, J. D. West, and C. R. Aragon. Toward the operationalization of visual metaphor. *J Assoc Info Sci Tech*, 2017: 68.
- [20] B. Keith and T. Mitra. Narrative Maps: An Algorithmic Approach to Represent and Extract Information Narratives. 2020. *ArXiv*, abs/2009.04508.
- [21] T. Xu, P. Zhang, Q. Huang, H. Zhang, Z. Gan, X. Huang and X. He. AttnGAN: Fine-Grained Text to Image Generation with Attentional Generative Adversarial Networks. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, 2017:1316-1324.
- [22] H. Sarma, R. Porzel, J.D., Smeddinck, R. Malaka and A. B. Samaddar. A Text to Animation System for Physical Exercises. *The Comp J*, 2018; 61: 589-1604.
- [23] A. X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su, J. Xiao, L. Yi and F. Yu. ShapeNet: An Information-Rich 3D Model Repository. In *Proceedings of CVPR*, 2015. *arXiv*: 1512.03012..
- [24] R. Ma, A. Patil, M. Fisher, M. Li, S. Pirk, B. Hua, S. Yeung, X. Tong, L. Guibas and H. Zhang. Language-driven synthesis of 3D scenes from scene databases. *ACM Trans Graph (TOG)* 2018; 37:1 – 16.
- [25] Y. Li, M. R. Min, D. Shen, D. E. Carlson and L. Carin. Video generation from text. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, New Orleans, Louisiana, USA, February 2-7, 2017.
- [26] B. Coyne and R. Sproat. WordsEye: An Automatic Text-to-Scene Conversion System" In *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques*, August 12, 2001, pp. 487-496.
- [27] C. L. Zitnick, D. Parikh and L. Vanderwende. Learning the Visual Interpretation of sentences. In *Proc. ICCV*, 2013, pp 1681–1688.
- [28] A. X. Chang, A. Savva and C. D. Manning. Learning Spatial Knowledge for Text to 3D Scene Generation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
- [29] F. Tan, S. Feng, and V. Ordonez. Text2Scene: Generating Compositional Scenes from Textual Descriptions. *Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp 6703-6712.
- [30] F. Huq, N. Ahmed, A. Iqbal. Static and Animated 3D Scene Generation from Free-form Text Descriptions. *arXiv: Computer Vision and Pattern Recognition*, 2020.
- [31] J. Pustejovsky N. Krishnaswamy. Generating Simulations of Motion Events from Verbal Descriptions. *Proceedings of the Third Joint Conference on Lexical and Computational Semantics*, Dublin, Ireland, August 23-24, 2014, pp 99–109.
- [32] N. Krishnaswamy and J. Pustejovsky. VoxSim: A Visual Platform for Modeling Motion Language. *COLING*, 2016a
- [33] N. Krishnaswamy and J. Pustejovsky. Multimodal Semantic Simulations of Linguistically Underspecified Motion Events. *Spatial Cognition*, 2016b.
- [34] T. Gupta, D. Schwenk, A. Farhadi, D. Hoiem and A. Kembhavi. Imagine This! Scripts to Composition to Videos. In *Proceedings of 15<sup>th</sup> European Conference on Computer Vision (ECCV)*, Munich, Germany. 8-14 September, 2018, pp 610-626.
- [35] Y. Balaji, M. R. Min, B. Bai, R. Chellappa, R. and H. P. Graf. Conditional GAN with Discriminative Filter Generation for Text-to-Video Synthesis. *Proceedings of 28<sup>th</sup> International*, 2019.
- [36] J. Zhang, Y. Hao, L. Li, D. Sun and L. Yuan. StoryWizard: a framework for fast stylized story illustration. *The Visl Compr* 2012; 877-887.
- [37] K. Schwarz, P. Rojtborg, J. Caspar, I. Gurevych, M. Goesele and H. P. A. Lensch. Text-to-Video: Story Illustration from Online Photo Collections. In: *Setchi R., Jordanov I., Howlett R.J., Jain L.C. (eds) Knowledge-Based and Intelligent Information and Engineering Systems. KES 2010. Lecture Notes in Computer Science*, 6279. Springer, Berlin, Heidelberg, 2010.
- [38] T. Marwah, G. Mittal and V. N. Balasubramanian. Attentive semantic video generation using captions. *CoRR* abs/1708.05980, 2017.
- [39] D. Khurana, A. Koli, K. Khatter and S. Singh. Natural Language Processing: State of The Art, Current Trends and Challenges, *Computation and Language*, 2017. *arXiv*: 1708.05148.
- [40] M. Honnibal, I. Montani, S. V. Landeghem and A. Boyd. *SpaCy: Industrial-strength Natural Language Processing in Python*. Zenodo, 2020. doi:10.5281/zenodo.1212303.
- [41] S. Bird, L. Edward and K. Ewan. *Natural Language Processing with Python*. O'Reilly Media Inc. 2009.
- [42] C. D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard and D. McClosky. The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 2014. Pp 55-60.
- [43] M. Fisher, D. Ritchie, M. Savva, T. A. Funkhouser and P. Hanrahan. Example-based synthesis of 3D object arrangements. *ACM Transactions on Graphics (TOG)*. 2012; 31: 1 - 11.
- [44] A. Dai, A. X. Chang, M. Savva, M. Halber and T. A. Funkhouser. ScanNet: Richly-Annotated 3D Reconstructions of Indoor Scenes. *CVPR*, 2017; 5828-5839.
- [45] A. Handa, V. Patraucean, V. Badrinarayanan, S. Stent and R. Cipolla. SceneNet: Understanding real-world indoor scenes with synthetic data. *arXiv* preprint 2015. *arXiv*:1511.07041.
- [46] J. D. Choi, J. Tetreault and A. Stent. It depends: Dependency parser comparison using a web-based evaluation tool, In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics*. 2015.
- [47] K. Ortman, A. Roussel and S. Dipper. Evaluating Off-the-Shelf NLP Tools for German. *KONVENS*. 2019.
- [48] G. N. Santos. Compiled Experiments for Modern Physics. Physics Department, De La Salle University, Manila, 2015:1-71.
- [49] T. Eftimov, B. K. Seljak and P. Korošec. A Rule-Based Named-Entity Recognition Method for Knowledge Extraction of Evidence-Based Dietary Recommendations. *PLoS ONE* 2017:12.
- [50] G. Popovski, S. Kochev, B. Korousic-Seljak and T. Eftimov. FoodIE: A Rule-based Named-entity Recognition Method for Food Information Extraction. *ICPRAM*. 2019.
- [51] T. Al-Moslmi, M. G. Ocaña, A. L. Opdahl, and C. Veres. Named Entity Extraction for Knowledge Graphs: A Literature Overview. *IEEE Access*, 8, 2020: 32862-32881.
- [52] C. M. Dal CM, S. Freitas, P. R. Luzzardi, R. A. Cava, M. A. Winckler, M. S. Pimenta and L. P. Nedel. Evaluating Usability of Information Visualization Techniques. In *Proceedings of Brazilian Symposium on Human Factors in Computing Systems*, 2012: 40-51.
- [53] S. Dahlgren. Design and Evaluation of a Visualization Method in a GIS for Complex Electronic Warfare Assessment with Regards to Usability. *A Master's thesis*. Department of Computer and Information Science Linköping University. 2019.

# A Review on Artificial Intelligence in the Context of Industry 4.0

Shadi Banitaan<sup>1</sup>, Ghaith Al-refai<sup>2</sup>, Sattam Almatarneh<sup>3</sup>, Hebah Alquran<sup>4</sup>

ECECS Department

University of Detroit Mercy, Detroit, USA<sup>1</sup>

Department of Mechatronics and Artificial Intelligence

German Jordanian University, Amman, Jordan<sup>2</sup>

Department of Data Science and Artificial Intelligence

Zarqa University, Zarqa, Jordan<sup>3</sup>

School of Information Security and Applied Computing

Eastern Michigan University, Ypsilanti, USA<sup>4</sup>

**Abstract**—Artificial Intelligence (AI) is seen as the most promising among Industry 4.0 advancements for businesses. Artificial intelligence, defined as computer models that mimic intelligent behavior, is poised to unleash the next wave of digital disruption and bring a competitive advantage to the industry. The value of AI lies not in its models, but in the ways in which we can harness them. It is becoming more common for industry objects to be converted into intelligent objects that can sense, act, adapt, and behave in a given environment. Leaders in the industry will need to make deliberate choices about how, when, and where to deploy these technologies. Our work highlights some of the primary AI emerging trends in Industry 4.0. We also discuss the advantages, challenges, and applications of AI in Industry 4.0.

**Keywords**—Artificial intelligence; Industry 4.0; intelligent manufacturing; industry analysis

## I. INTRODUCTION

Industry 4.0 is a term used to describe the Fourth Industrial Revolution. Manufacturing technologies are part of this revolution. Among the technologies included in this group are technologies like the Internet of Things (IoT), cyber-physical systems (CPS), and artificial intelligence (AI). A machine's ability to perform human functions, such as learning, reasoning, and solving problems, is commonly referred to as artificial intelligence. Using sensor technologies, machine intelligence agents can perceive and interact with their surroundings.

Artificial Intelligence allows computer systems to learn from experience, adjust to new input data, and make intelligent tasks. Fig. 1 illustrates the major areas and technologies associated with artificial intelligence.

- Machine learning is the process of developing computer systems that can detect patterns from raw data. There are two major types of machine learning: supervised and unsupervised.
- Supervised learning involves algorithms that generate a predictive model from a set of training data, which includes both training observation/examples and labels [3].
- Unsupervised learning refers to the creation of a model from observations/examples that do not have class labels [3].

- Deep Learning is concerned with algorithms inspired by the structure and function of the brain called artificial neural networks. These algorithms are used for a variety of tasks, such as image classification, by learning from large amounts of data and using multiple layers of artificial neural networks to produce intelligent decisions [4].
- Natural Language Processing (NLP) is a subfield of computer science and artificial intelligence concerned with the interactions between computers and human languages, with a focus on making it possible for machines to read, understand, and generate human language [5].
- Expert systems are designed to mimic the decision-making abilities of a human expert in a specific domain. They use a combination of knowledge representation, inference rules, and a database of facts to provide reasoning and advice to solve complex problems in fields such as medicine, finance, and engineering [6].
- Computer Vision is concerned with enabling computers to interpret and understand visual information from the world in the same way that humans do. It involves the development of models to perform tasks such as image and video recognition, object detection, and image segmentation [7].
- Speech Recognition, also known as Automatic Speech Recognition (ASR), is concerned with the ability of computers to recognize, understand, and transcribe human speech. The goal of speech recognition is to develop algorithms that can accurately transcribe or translate spoken language in real-time, enabling natural and convenient human-computer interaction [8].

AI is considered the next revolution in health care, manufacturing, and mobility. AI plays a vital role in Intelligent Manufacturing Systems (IMS) by introducing learning, acting, and reasoning. Manufacturing objects are transformed into intelligent objects that can self-correct without human intervention [11], [60]. Manufacturing will benefit from AI if it is able to harness new capabilities, many of which have seen

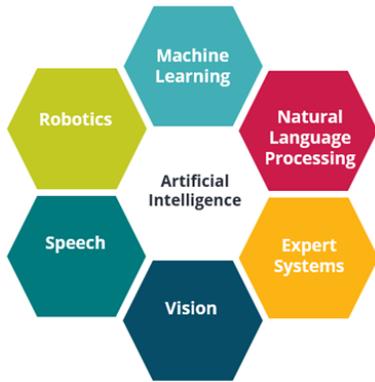


Fig. 1. Key AI areas and tech

dramatic increases in recent years [2].

AI combined with emerging technologies such as Big Data, Blockchain, and IoT can eliminate downtime, maximize throughput, and improve efficiencies. For example, FANUC and Cisco have optimized systems that enhance value for manufacturers [18], [47]. To reach the goal of autonomous machines in Industry 4.0, blockchain can be utilized by connecting the ERP, parts supplier, and the cyber-physical system in a factory, enabling the machines to order replacement parts securely and independently. Additionally, blockchain's ability to facilitate seamless and transparent financial transactions between smart devices is essential for the economic changes brought by Industry 4.0 [10].

The Reference Architecture Model Industry 4.0 defines the Industry 4.0 layers [16]. It consists of the following layers [16]:

- The business layer organizes business operations and connections between different processes, adhering to the legal and regulatory restrictions, to support the underlying business models.
- The functional layer describes an asset's role in Industry 4.0-based systems based on its logical functions.
- The information layer represents the assets' technical features, like services and data.
- The communication layer provides regular communication between the integration layer and the information layer about services and data.
- The integration layer represents the physical assets, and the digital capability provides computer-aided control and creates events based on the assets.
- The assets layer serves the physical world, such as physical objects, software, and actors in the physical world.

The service-oriented RAMI 4.0 goes deeper into representing digital manufacturing models [16].

LinkedIn reported in 2018 that six of the top 15 emerging jobs were related to AI, with positions requiring expertise in deep learning experiencing the highest growth, according to data from Monster.com [22]. Deep Learning is a branch of

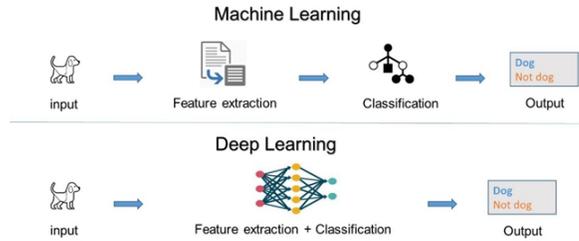


Fig. 2. Comparison between machine learning and deep learning

machine learning that utilizes multiple non-linear layers for feature extraction, transformation, and classification, either in a supervised or unsupervised manner [23].

Deep learning and classical machine learning are intended to model the relationship between inputs and outputs. Deep Learning distinguishes itself from traditional machine learning in its approach to feature learning, model building, and training. It combines these elements into one model, adjusting kernels or tuning parameters for optimal results [26]. Fig. 2 shows the main difference between traditional machine learning and deep learning.

Deep learning revolutionizes manufacturing by transforming facilities into highly efficient smart operations, leading to reduced operating costs, increased productivity, and decreased downtime [26]. Fig. 5 illustrates the main differences between machine learning and deep learning. Deep learning eliminates the need for expert involvement by incrementally learning high-level features from data, while traditional machine learning requires domain experts to identify features.

The remaining sections are organized as follows. Section II describes the methodology we followed in conducting the review. Section III presents the typical applications of AI in industry and shows some use cases. Section IV discusses the advantages and the challenges that are currently noticeable by using AI. Section V demonstrates the industry analysis. Section VI reviews some future trends. Section VII concludes the paper.

## II. METHODOLOGY

This study follows the two-stage approach developed by Webster and Watson[9] for reviewing relevant literature. As a first step, the following search phrases were used to search for papers published between 2016 and 2020 on Google Scholar and ScienceDirect:

- "Industry 4.0" & "Artificial Intelligence"
- "Industry 4.0" & "Trends"
- "Industry 4.0" & "Intelligent Manufacturing"

The search returned 176 results. In the second step, these papers were carefully reviewed, and unrelated papers were eliminated. We have compiled a final list of 39 relevant articles. Fig. 3 and 4 show the publication years and citation numbers of these papers. For example, 13 relevant articles have a total of 1400 citations in 2018. The selected papers are then grouped into four research categories, as shown in Table I. The

TABLE I. RESEARCH CATEGORIES OF THE SELECTED PUBLICATIONS

Research Category	No. of Publications
Applications of AI in Industry 4.0	24
AI technologies and approaches in the Industry 4.0	10
Advantages and Challenges	3
Emerging Trends	2

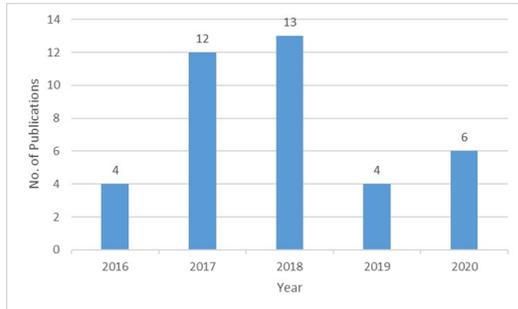


Fig. 3. The number of publications per year (2016–2020)

categories’ distribution shows that more attention has been paid to AI applications in Industry 4.0, followed by the primary approaches and methods of AI in Industry 4.0. Fig. 5 shows the paper organization’s block diagram and an overview of AI in Industry 4.0.

### III. AI APPLICATIONS AND USE CASES

This section represents applications of AI in major industries. Some case studies are then presented.

#### A. Applications

Today’s society uses artificial intelligence in a variety of ways. It has been used to develop and advance many fields and industries, including aerospace, automotive, electronics, finance, medical, education, retail, and more. Ahuett-Garza et al. provided a brief review of machine learning, IoT, and adaptive manufacturing in industry 4 [56]. Preuveneers et al. developed a study in AI and machine learning in intelligent manufacturing environments settings [57]

Intelligent manufacturing tools and models are explained in [58]. Lee et al. proposed that industrial AI’s main elements

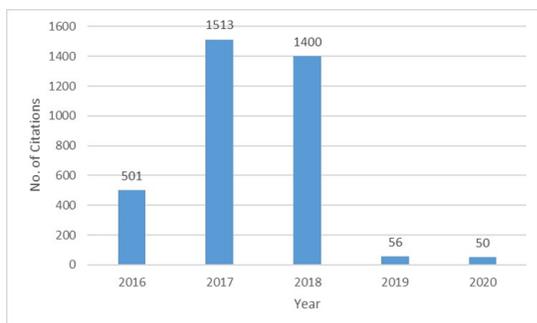


Fig. 4. The number of citations per year (2016–2020)

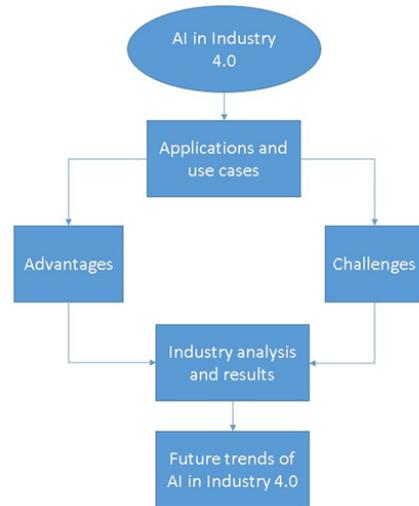


Fig. 5. The organization of the paper

include analytics technology, cyber technology, and big data technology [59]. Cheng et al. discussed the future development direction of Industry 4.0, which provides a reference for its intelligent manufacturing [61]. Liu et al. presented the manufacturing demonstration system based on IoT in Industry 4.0 [62]. Table II shows some main applications of AI in several industrial sectors.

#### B. Use Cases

This section includes several examples of the successful implementation of AI in Industry 4.0.

1) *Doxel robots use AI to improve accuracy and efficiency on large construction projects:* A new robot can check that building projects are going planned using AI and LIDAR. After a construction site shuts down for the day, robots created by Doxel can start working. Using LIDAR, it scans the construction sites and feeds that data into deep-learning algorithms.

The deep learning algorithms find anything that deviates from building plans so that the management team can fix the problems the next day. The main premise is that if errors have not been noticed directly on the worksite, they will create complex issues that take time and money to fix. Instant problem resolution leads to substantial cost savings. A recent test of the approach on an office building project improved labor productivity by around 38%.

2) *Anomaly detection of bearings at Altair engineering:* Bearings play a crucial role in the automotive sector. This example uses sensor data from four bearings, sampled at 20kHz, resulting in 1-second sampling every 10 minutes for 9 days. The dataset originates from NASA’s Prognostics Center of Excellence. The initial sampling, representing a new bearing, serves as a reference for detecting anomalies. The goal is to monitor the bearings’ health as they age and predict the beginning of degradation, which will be flagged as an anomaly to the user. Identifying anomalies enables the user to plan

TABLE II. APPLICATIONS OF AI IN INDUSTRY

Industry	Applications	Summary	Technology/Technique	Ref.
Aerospace	- Fuel consumption prediction	- A genetic algorithm-optimized neural network topology is designed to predict the fuel flow-rate of a transport aircraft	- Feed-forward backpropagation, Levenberg–Marquardt algorithms, and genetic algorithms	[25]
	- Aircraft failure times prediction	- Predict when the failure will happen by aircraft type and age	- Artificial neural networks and genetic algorithms	[15]
	- Aircraft design cycle time reduction	- AI is used to expedite the decision-making process in the early stages of the aircraft design process	- fuzzy logic and neural network	[14]
Automotive	- Driving Assistance and Autonomous Driving	- State-of-the-art deep learning technologies used in autonomous driving	- AI-based self-driving architectures, convolutional and recurrent neural networks, and reinforcement learning	[17], [37]
	- Driver monitoring	- Monitor drivers and identify driving tasks in vehicles	- Kinect, Random Forest, and Feedforward Neural Network	[13]
	- Vehicle manufacturing	- Human–collaborative robot assembly in cyber-physical production. Manufacturing system produces products from scratch without any human intervention during the process	- collaborative robots and additive manufacturing	[12, 53]
Electronics	- Diagnosis of electrical machines and drives	The AI methodologies are applied to an induction machine, utilizing as input signals the instantaneous voltages and currents	- Expert systems, artificial neural networks, and fuzzy logic systems	[50]
	- Smart refrigerators	- The intelligent refrigerator is capable of sensing and monitoring its contents and notifying the user about the type, quantity, and freshness of the food in the refrigerator	- Internet of Things, Mobile Internet, advanced control and sensing, and food preserving technology	[20], [30]
	- Intelligent video surveillance systems	- Intelligent video surveillance systems that control private and public places and detect dangerous situations	- Neural networks, classification, and clustering techniques	[24]
Medical	- Breast cancer detection	- Performance comparison between different machine learning algorithms on the Wisconsin Breast Cancer datasets	- Support Vector Machine, Decision Tree, Naive Bayes, and k Nearest Neighbors	[39]
	- Disease Diagnosis and symptoms checker	- Artificial intelligence platform for self-diagnosis and symptom checker	- Artificial Intelligence	[40]
	- Robot-assisted AI surgery	- deep learning approach for robotic machine segmentation	- Deep neural network	[43]
Retail	- Customer segmentation and clustering	- Systematic Approach to Customer Segmentation and Buyer Targeting for Profit Maximization	- K-mean clustering	[44]
	- Retail recommendation engines using machine learning	- Use of Deep Learning in Modern Recommendation System: A Summary of Recent Works	- Natural language processing, Deep Learning	[45]
	- RFID network planning	- Artificial intelligent paradigm is developed to model and optimize RFID networks	- Artificial neural networks and computational artificial intelligence algorithms	[46]
Telecom	- Smart traffic prediction and path optimization	- Optimizing space-air-ground integrated networks by artificial Intelligence	- Deep Learning	[49]
	- Security	- Articulating a comprehensive threat model for ML and categorize attacks and defenses within an adversarial framework	- Machine learning	[51]
	- Customer segmentation based on mobile usage	- Identifying prime customers based on mobile usage patterns	- Support vector machine and K-mean clustering	[52]

proper maintenance of the bearings, avoiding potential failures and irreversible issues.

Initially, Principal Component Analysis (PCA) is applied for dimension reduction. The samples are then compared to the healthy sample to assess the current health of the bearing, and the comparison is represented as a Health Index (HI). An anomaly is detected if there is a 95% decrease in correlation in 5 or more out of 10 consecutive samples. The entire machine learning model is located on an edge device and transmits the HI and anomaly status of each vibration pattern from the sensor in real-time to Altair’s SmartSight. The user can view the status graphically and, if an anomaly is detected, Altair’s SmartCore sends an email alert to the user.

3) *AI from the factory floor to the showroom at Mercedes-Benz:* The widespread use of AI in the automotive industry’s manufacturing process is well documented. OEMs are now incorporating AI into all areas of their business, including sales. With AI insights, companies can determine the best product segment to sell, to whom, and when. Mercedes-Benz, a large-scale truck and bus production plant in Brazil, uses Azure Machine Learning to transform its sales process. The tool combines internal and external data, such as registration numbers, macroeconomic indicators, local laws, sales info, and stats, to aid the brand’s sales reps in making tailored offers at the right time. The system improves with each monthly data report inputted by dealers, resulting in more accurate recommendations.

4) *AI from the ford motor company and Argo AI:* Ford Motor Company announced a partnership with Argo AI in 2017, investing \$1 billion in the virtual driver system for its SAE Level 4 autonomous vehicles [36]. The vehicles, equipped with Argo’s cutting-edge machine learning and computer vision technology, will be deployed for ride-hailing and delivery services in several cities in the US.

#### IV. ADVANTAGES AND CHALLENGES

This section highlights the significant advantages and challenges currently noticeable in using AI, especially in industrial environments.

##### A. Advantages

Significant savings in labor costs due to troubleshooting, maintenance, and repair: AI and machine learning allow machines and computers to replace human labor in many tasks, such as manufacturing, agriculture, and business services. Acemoglu et al. discussed the impact of AI in various industries and their economic impacts [28]. Business decision-makers realize that AI can help create new products, services, and business models [31], [35].

Improved reliability and efficiency through extended time between failures: AI enhances systems’ efficiency in various ways, including more accurate demand forecasting to better



Fig. 6. How companies are adopting AI

manage goods inventory and storage. AI's predictive maintenance helps prevent costly, unexpected machine shutdowns and maintenance in factories [55]. Many examples of AI's ability to improve systems efficiency have been discussed by Jimenez et al. [32].

Safer work environments through AI's ability to detect and respond to hazardous situations: Workplace injuries can be costly for businesses. AI can perform dangerous tasks, such as in construction, heavy machinery manufacturing, and oil and gas plants, reducing the risk of injury to workers. An example of AI enhancing workplace safety is provided in [33].

AI will generate new employment opportunities in various sectors, including cybersecurity, data analysis, machine learning, deep learning algorithms, and data science.

### B. Challenges

Implementing AI comes with its own set of challenges and problems. The issues to be aware of are outlined in Table III, along with recommended solutions.

## V. INDUSTRY ANALYSIS

The results of a survey conducted by Vanson Bourne in July 2017, with 260 respondents, showed that the main obstacles to the implementation of AI are a lack of IT infrastructure and a shortage of talent, as depicted in Fig. 6 [34]. The development of mature AI infrastructure is likely as the world moves towards IoT, smart cities, and cloud systems. The shortage of talent in the field of AI can be addressed through the creation of graduate certificates and programs focused on AI and machine learning offered by universities.

A survey by McKinsey Global Institute found that AI implementation outside of the tech sector is still in its early stages. Only 20% of 3,000 executives from ten countries and 14 industries reported using AI technology in a significant or central aspect of their business. Many companies are uncertain about the potential benefits and return on investment from AI. The study analyzed over 160 use cases and found that only 12% of these employed AI commercially [19]. Fig. 7 illustrates how companies are implementing AI.

A survey of Michigan-based small and medium-sized manufacturing executives conducted by Automation Alley in 2019 showed that only 22% of the companies were currently using AI, 37% were planning to implement it in the next year, and 4% had no plans to use AI (as shown in Fig. 7). The adoption of AI is expected to drive growth and improve revenue for companies that implement it, while those that don't adopt it risk falling behind.

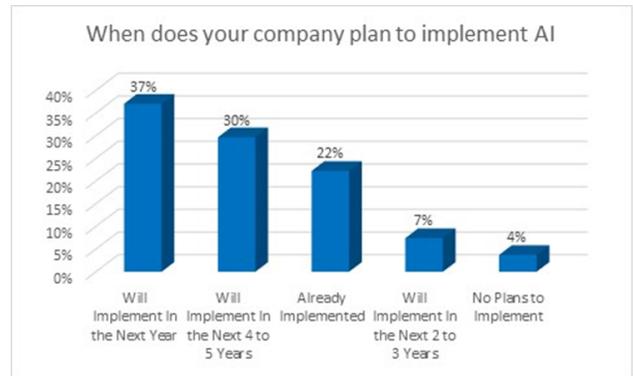


Fig. 7. When does your company plan to implement AI

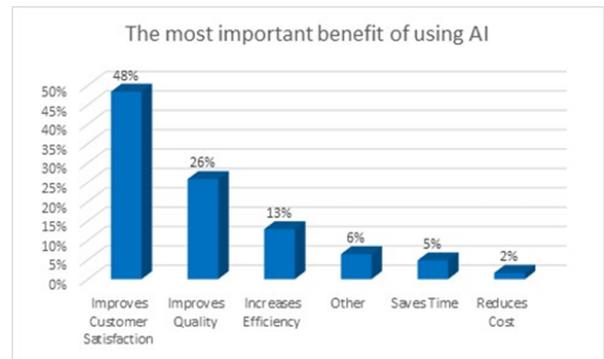


Fig. 8. The most important benefit of using AI

Fig. 8 shows what those Michigan executives surveyed believe to be the essential benefit of using AI, including 1) Improving customer satisfaction, 2) Improving quality, and 3) Increasing efficiency. Initial training and lack of understanding by senior management are the biggest challenges of using AI based on Automation Alley's survey results, as shown in Fig. 9. Many AI consultation companies can provide training and frameworks for AI systems in various business sectors. We believe that identifying AI barriers can lead to accelerating AI adoption.

AI implementation in a business requires the higher management to have the flexibility to change, openness and vision,

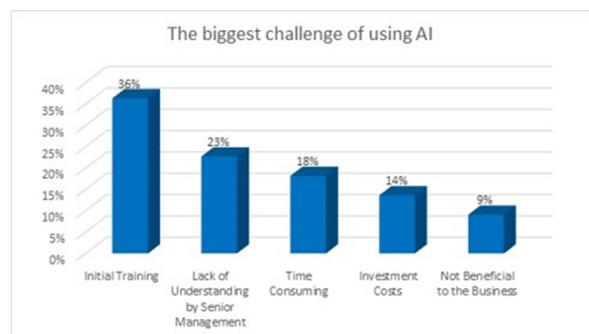


Fig. 9. The biggest challenge of using AI

TABLE III. AI CHALLENGES

Challenge	Description	Suggested Solution	Ref.
Quantity of data	One of the major challenges in manufacturing systems is handling and utilizing the large amounts of data generated, including understanding, cleaning, using, and storing it.	The suggested solution incorporates the latest cloud computing and high-performance cloud techniques to create a large-scale remote sensing data processing system that can handle on-demand real-time services	[38]
Privacy and ownership issues	The widespread collection of data by corporations raises privacy concerns, leading society to consider questions of data collection scope, accessibility, and ownership	Strong privacy guarantees, competitive advantages, and discriminatory policies	[29]
Difficulty collecting and labeling data	Collecting and labeling data is a costly, time consuming, and lengthy process	Transfer learning can be used to tackle this problem by creating high-performance learners that are trained with easily obtained data from different domains	[41]
Difficulty generalizing the results	It is hard to find one algorithm to be effective across a range of inputs and applications	The proposed learning theory analyzes individual problem instances using measure theory instead of sets of instances using statistics. It is not meant to compete with previous learning theories, but rather to complement them, due to differences in assumptions and objectives	[42]
Computation and run-time for AI algorithms	Machine learning and deep learning algorithms require heavy computation and large memory to perform in real-time applications	Algorithms optimization, virtual machines, and cloud computation	[54]

long-term objectives, and effective collaboration. However, there are also some specific challenges where executives may still need to learn more about AI and increase their knowledge on how to organize their business using AI [48].

## VI. EMERGING TRENDS

The major AI trends are discussed in the following subsections.

### A. Smart Devices

The manufacturing industry necessitates the use of sophisticated smart digital devices. Industry 4.0 utilizes AI and IoT to create intelligent objects. Future AI and IoT will possess features like self-configuration, self-defense, self-repair, and self-improvement [1].

### B. Manufacturing Systems

A predictive manufacturing system (PMS) is an intelligent manufacturing system that provides self-awareness, self-predictability, self-maintenance, and self-learning capabilities. In PMS, various technologies and techniques, such as statistics, data mining, models, and AI methods, are used to convert data into information, identify uncertainties, and make predictions about manufacturing systems [21].

The PMS conceptual framework includes a platform, predictive analytics, and visualization tools. Data is generated by the monitored assets. Platforms are chosen based on several factors, including computation speed and investment cost. The purpose of predictive analytics is to extract and predict future outcomes and trends. Among the benefits of PMS are cost reductions, improved operation efficiency, and improved product quality.

### C. Human-Machine Interaction and AI

The integration of AI with human-machine interaction is a crucial aspect in constructing Industry 4.0 enterprises. The goal is to optimize efficiency by seamlessly connecting human dynamics with hardware and software in machine-to-human and human-to-machine interfaces. One example is the AI-powered workspace where humans and machines work together to achieve outcomes not possible by either alone [27].

## VII. CONCLUSION

The meaning of AI and its subfields are continuously evolving. Industry 4.0's IoT, Big Data, cloud and cybersecurity pave the path for AI implementation and usage. Deep Learning revolutionizes manufacturing into highly efficient smart facilities. IT infrastructure investment and a talent pipeline for AI are crucial for unlocking its potential. The most common benefits of AI are improving customer value and quality. Adopters of AI should be aware of data privacy and ownership challenges, the cost of labeling data, and difficulties in generalizing results. AI adoption outside the technology sector is still in its early stages and manufacturers are aware of AI but its use is still limited. Future developments will involve further exploration of AI and its implementation in industrial settings.

## REFERENCES

- [1] Xu, Li Da, Eric L. Xu, and Ling Li." Industry 4.0: state of the art and future trends." International Journal of Production Research 56, no. 8 (2018): 2941-2962.
- [2] Banjanovi'c-Mehmedovi'c, L., Mehmedovi'c, F. (2020). Intelligent Manufacturing Systems Driven by Artificial Intelligence in Industry 4.0. Handbook of Research on Integrating Industry 4.0 in Business and Manufacturing, 31-52.
- [3] Han, J., Pei, J., Kamber, M. (2011). Data mining: concepts and techniques. Elsevier.
- [4] Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. Deep Learning. MIT press, 2016.
- [5] Jackson, Peter, and Isabelle Moulinier. Natural language processing for online applications: Text retrieval, extraction and categorization. Vol. 5. John Benjamins Publishing, 2007.
- [6] Tan, Haocheng." A brief history and technical review of the expert system research." In IOP Conference Series: Materials Science and Engineering, vol. 242, no. 1, p. 012111. IOP Publishing, 2017.
- [7] Nixon, Mark, and Alberto Aguado. Feature extraction and image processing for computer vision. Academic Press, 2019.
- [8] Bangalore, Srinivas, Robert Bell, Diamantino Antonio Caseiro, Mazin Gilbert, and Patrick Haffner." System and method for rapid customization of speech recognition models." U.S. Patent 9,679,561, issued June 13, 2017.
- [9] Webster, J., Watson, R. T. (2002). Analyzing the past to prepare for the future: Writing a literature review. MIS quarterly, xiii-xxiii.
- [10] Joshi, N. 2017. "Blockchain Meets Industry 4.0-What Happened Next?" <https://www.allerin.com/blog/5659-2>
- [11] Zhong, Ray Y., Xun Xu, Eberhard Klotz, and Stephen T. Newman." Intelligent manufacturing in the context of industry 4.0: a review." Engineering 3, no. 5 (2017): 616-630.

- [12] Wang, Xi Vincent, Zsolt Kem'eny, Jo'zsef Va'nca, and Lihui Wang." Human-robot collaborative assembly in cyber-physical production: Classification framework and implementation." *CIRP annals* 66, no. 1 (2017): 5-8.
- [13] Xing, Yang, Chen Lv, Zhaozhong Zhang, Huaji Wang, Xiaoxiang Na, Dongpu Cao, Efstathios Velenis, and Fei-Yue Wang." Identification and analysis of driver postures for in-vehicle driving activities and secondary tasks recognition." *IEEE Transactions on Computational Social Systems* 5, no. 1 (2017): 95-108.
- [14] Oroumieh, Mehran Ali Azizi, S. Mohammad Bagher Malaek, Mahmud Ashrafzaadeh, and S. Mahmoud Taheri." Aircraft design cycle time reduction using artificial Intelligence." *Aerospace science and technology* 26, no. 1 (2013): 244-258.
- [15] Altay, Ayca, Omer Ozkan, and Gulgun Kayakutlu." Prediction of aircraft failure times using artificial neural networks and genetic algorithms." *Journal of Aircraft* 51, no. 1 (2014): 47-53.
- [16] Lin, S. W., Murphy, B., Clauer, E., Loewen, U., Neubert, R., Bachmann, G., ... Hankel, M. (2017). *Architecture Alignment and Interoperability-An Industrial Internet Consortium and Plattform Industrie 4.0 Joint Whitepaper*. White Paper, Industrial Internet Consortium.
- [17] Grigorescu, S., Trasnea, B., Cocias, T., Macesanu, G. (2019). A survey of deep learning techniques for autonomous driving. *Journal of Field Robotics*.
- [18] Green, T. Zero Down Time: FANUC Explains It All at RoboBusiness Review. Boston, MA. <https://www.roboticsbusinessreview.com/manufacturing/zero-time-fanuc-explains-robobusiness/> (2016)
- [19] Chui, M., and Francisco, S. (2017). *Artificial Intelligence the next digital frontier?*. McKinsey and Company Global Institute, 47.
- [20] Shweta, A. S." Intelligent refrigerator using ARTIFICIAL INTELLIGENCE." In 2017 11th International Conference on Intelligent Systems and Control (ISCO), pp. 464-468. IEEE, 2017.
- [21] Nikolic, Bojana, Jelena Ignjatovic, Nikola Suzic, Branislav Stevanov, and Aleksandar Rikalovic." PREDICTIVE MANUFACTURING SYSTEMS IN INDUSTRY 4.0: TRENDS, BENEFITS AND CHALLENGES." *Annals of DAAAM and Proceedings* 28 (2017).
- [22] Press, Gil." Top 10 hot artificial intelligence (AI) technologies." *Forbes*, viewed 23 (2017).
- [23] Deng, L., and Yu, D. (2014). *Deep Learning: methods and applications*. Foundations and Trends in Signal Processing, 7(3-4), 197-387.
- [24] Mabrouk, Amira Ben, and Ezzeddine Zagrouba." Abnormal behavior recognition for intelligent video surveillance systems: A review." *Expert Systems with Applications* 91 (2018): 480-491.
- [25] Baklacioglu, Tolga." Modeling the fuel flow-rate of transport aircraft during flight phases using genetic algorithm-optimized neural networks." *Aerospace Science and Technology* 49 (2016): 52-62.
- [26] Wang, J., Ma, Y., Zhang, L., Gao, R. X., and Wu, D. (2018). *Deep Learning for smart manufacturing: Methods and applications*. *Journal of Manufacturing Systems*, 48, 144-156.
- [27] Schaeffer, E. (2017). *Industry X.0 Realizing Digital Value in Industrial Sectors*. Redline Verlag, Munich, Germany.
- [28] Acemoglu, Daron, and Pascual Restrepo. *Artificial Intelligence, automation and work*. No. w24196. National Bureau of Economic Research, 2018.
- [29] Young, Meg, Luke Rodriguez, Emily Keller, Feiyang Sun, Boyang Sa, Jan Whittington, and Bill Howe." Beyond open vs. closed: Balancing individual privacy and public accountability in data sharing." In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 191-200. 2019.
- [30] Gao, Xiaoyan, Xiangqian Ding, Ruichun Hou, and Ye Tao." Research on Food Recognition of Smart Refrigerator Based on SSD Target Detection Algorithm." In *Proceedings of the 2019 International Conference on Artificial Intelligence and Computer Science*, pp. 303-308. 2019.
- [31] McKendrick, J. Artificial Intelligence Doesn't Just Cut Costs, It Expands Business Brainpower. <https://www.forbes.com/sites/joemckendrick/2017/01/24/artificial-intelligence-doesnt-just-cut-costs-it-expands-business-brainpower/65f7e73b535a> (2017)
- [32] Jimenez, J." 5 Ways Artificial Intelligence Can Boost Productivity." Internet: <https://www.industryweek.com/technology-and-iiot/5-ways-artificial-intelligence-can-boost-productivity> (2018)
- [33] Innovation potential, How AI is Making Workers Safe, <https://www.apexofinnovation.com/how-ai-is-making-workers-safe> (2019)
- [34] Teradata, State of Artificial Intelligence for Enterprises, [Online] Available at: <http://assets.teradata.com/resourceCenter/downloads/AnalystReports/TeradataReport>
- [35] Chui, M., Manyika, J., Miremadi, M., Henke, N., Chung, R., Nel, P., and Malhotra, S. (2018). *Notes from the AI frontier: Insights from hundreds of use cases*.
- [36] Shekhar, S. (2017, February 13). *Ford Invests \$1 Billion in Argo AI for Autonomous Vehicle Leadership*. PC Quest.
- [37] Sallab, A. E., Abdou, M., Perot, E., Yogamani, S. (2017). *Deep reinforcement learning framework for autonomous driving*. *Electronic Imaging*, 2017(19), 70-76.
- [38] Wang, Lizhe, Yan Ma, Jining Yan, Victor Chang, and Albert Y. Zomaya." pipsCloud: High performance cloud computing for remote sensing big data management and processing." *Future Generation Computer Systems* 78 (2018): 353-368.
- [39] Asri, Hiba, Hajar Mousannif, Hassan Al Moatassime, and Thomas Noel." Using machine learning algorithms for breast cancer risk prediction and diagnosis." *Procedia Computer Science* 83 (2016): 1064-1069.
- [40] Razzaki, Salman, Adam Baker, Yura Perov, Katherine Middleton, Janie Baxter, Daniel Mullarkey, Davinder Sangar et al." A comparative study of artificial Intelligence and human doctors for the purpose of triage and diagnosis." *arXiv preprint arXiv:1806.10698* (2018).
- [41] Weiss, Karl, Taghi M. Khoshgoftaar, and DingDing Wang." A survey of transfer learning." *Journal of Big data* 3, no. 1 (2016): 9.
- [42] Kawaguchi, Kenji, Yoshua Bengio, Vikas Verma, and Leslie Pack Kaelbling." Generalization in machine learning via analytical learning theory." *stat* 1050 (2019): 6.
- [43] Shvets, Alexey A., Alexander Rakhlin, Alexandr A. Kalinin, and Vladimir I. Iglovikov." Automatic instrument segmentation in robot-assisted surgery using deep learning." In 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA), pp. 624-628. IEEE, 2018.
- [44] Bhade, Kalyani, Vedanti Gulalkari, Nidhi Harwani, and Sudhir N. Dhage." A Systematic Approach to Customer Segmentation and Buyer Targeting for Profit Maximization." In 2018 9th International Conference on Computing, Communication and Networking Technologies (ICCCNT), pp. 1-6. IEEE, 2018.
- [45] Singhal, Ayush, Pradeep Sinha, and Rakesh Pant." use of deep learning in modern recommendation system: A summary of recent works." *arXiv preprint arXiv:1712.07525* (2017).
- [46] Azizi, A. (2019). *Applications of Artificial Intelligence Techniques in Industry 4.0*. Berlin, Germany: Springer.
- [47] Daki, Houda, Asmaa El Hannani, Abdelhak Aqqal, Abdelfattah Haidine, and Aziz Dahbi." Big Data management in smart grid: concepts, requirements and implementation." *Journal of Big Data* 4, no. 1 (2017): 1-19.
- [48] Ransbotham, Sam, David Kiron, Philipp Gerbert, and Martin Reeves." Reshaping business with artificial Intelligence: Closing the gap between ambition and action." *MIT Sloan Management Review* 59, no. 1 (2017)
- [49] Kato, Nei, Zubair Md Fadlullah, Fengxiao Tang, Bomin Mao, Shigenori Tani, Atsushi Okamura, and Jijia Liu." Optimizing space-air-ground integrated networks by artificial Intelligence." *IEEE Wireless Communications* 26, no. 4 (2019): 140-147.
- [50] Filippetti, Fiorenzo, Giovanni Franceschini, Carla Tassoni, and Peter Vas." Recent developments of induction motor drives fault diagnosis using AI techniques." *IEEE transactions on industrial electronics* 47, no. 5 (2000): 994-1004.
- [51] Papernot, Nicolas, Patrick McDaniel, Arunesh Sinha, and Michael Wellman." Towards the science of security and privacy in machine learning." *arXiv preprint arXiv:1611.03814* (2016).
- [52] Arora, Deepali, and Kin Fun Li." Identifying Prime Customers Based on Mobile Usage Patterns." In *International Conference on P2P, Parallel, Grid, Cloud and Internet Computing*, pp. 853-861. Springer, Cham, 2016.

- [53] Kaleem, M. A., Khan, M. (2020) Significance of Additive Manufacturing for Industry 4.0 With Introduction of Artificial Intelligence in Additive Manufacturing Regimes.
- [54] Curtis, Frank E., and Katya Scheinberg." Optimization methods for supervised machine learning: From linear models to deep learning." In *Leading Developments from INFORMS Communities*, pp. 89-114. INFORMS, 2017.
- [55] Martinez, L. R., Rios, R. A. O., Prieto, M. D. (2020). New Trends in the Use of Artificial Intelligence for the Industry 4.0.
- [56] Ahuett-Garza, Horacio, and Thomas Kurfess." A brief discussion on the trends of habilitating technologies for Industry 4.0 and Smart manufacturing." *Manufacturing Letters* 15 (2018): 60-63.
- [57] Preuveneers, Davy, and Elisabeth Ilie-Zudor." The intelligent industry of the future: A survey on emerging trends, research challenges and opportunities in Industry 4.0." *Journal of Ambient Intelligence and Smart Environments* 9, no. 3 (2017): 287-298.
- [58] Yao, Xifan, Jiajun Zhou, Jiangming Zhang, and Claudio R. Bo"er." From intelligent manufacturing to smart manufacturing for industry 4.0 driven by next generation artificial intelligence and further on." In *2017 5th international conference on enterprise systems (ES)*, pp. 311-318. IEEE, 2017.
- [59] Lee, Jay, Hossein Davari, Jaskaran Singh, and Vibhor Pandhare." Industrial Artificial Intelligence for industry 4.0-based manufacturing systems." *Manufacturing letters* 18 (2018): 20-23.
- [60] Li, B. H., Hou, B. C., Yu, W. T., Lu, X. B., Yang, C. W. (2017). Applications of Artificial Intelligence in intelligent manufacturing: a review. *Frontiers of Information Technology Electronic Engineering*, 18(1), 86-96.
- [61] Cheng, Guo-Jian, Li-Ting Liu, Xin-Jian Qiang, and Ye Liu." Industry 4.0 development and application of intelligent manufacturing." In *2016 international conference on information system and artificial intelligence (ISAI)*, pp. 407-410. IEEE, 2016.
- [62] Liu, Y., Li, Z., Wang, Z., Bai, H., Xing, Y., Zeng, P. (2019, April). Design of the intelligent manufacturing demonstration system based on IoT in the context of industry 4.0. In *IOP Conference Series: Earth and Environmental Science* (Vol. 252, No. 5, p. 052001). IOP Publishing.

# A Machine Learning Enabled Hall-Effect IoT-System for Monitoring Building Vibrations

Emanuele Lattanzi, Paolo Capellacci, Valerio Freschi  
Dep. of Pure and Applied Sciences, University of Urbino,  
Italy

**Abstract**—Vibration monitoring of civil infrastructures is a fundamental task to assess their structural health, which can be nowadays carried on at reduced costs thanks to new sensing devices and embedded hardware platforms. In this work, we present a system for monitoring vibrations in buildings based on a novel, cheap, Hall-effect vibration sensor that is interfaced with a commercially available embedded hardware platform, in order to support communication toward cloud based services by means of IoT communication protocols. Two deep learning neural networks have been implemented and tested to demonstrate the capability of performing nontrivial prediction tasks directly on board of the embedded platform, an important feature to conceive dynamical policies for deciding whether to perform a recognition task on the final (resource constrained) device, or delegate it to the cloud according to specific energy, latency, accuracy requirements. Experimental evaluation on two use cases, namely the detection of a seismic event and the count of steps made by people transiting in a public building highlight the potential of the adopted solution; for instance, recognition of walking-induced vibrations can be achieved with an accuracy of 96% in real-time within time windows of 500ms. Overall, the results of the empirical investigation show the flexibility of the proposed solution as a promising alternative for the design of vibration monitoring systems in built environments.

**Keywords**—Vibration sensor; machine learning; hall-effect

## I. INTRODUCTION

The growing attention toward the structural efficiency of civil structures and infrastructures prompts the need for the design of novel and effective systems capable of monitoring structural health at adequate time and space resolution, and providing effective evaluation and support for downstream decision-making in order to mitigate risks, increase safety, reduce maintenance costs [1], [2].

Structural Health Monitoring (SHM), i.e. the set of techniques, methodologies, and technologies that enable to obtain information regarding the state of a structure (for instance in terms of the functioning of structures or their response), is therefore becoming a crucial aspect of the lifetime cycle of the built environment.

The analysis of the vibrational response of a building, encoded into a time series of displacements and/or of accelerations is a commonly adopted approach for SHM, for instance for establishing potential damage states. This response is usually the result of the non-trivial interplay between several physical properties of the structure (e.g. mass distribution, stiffness, damage pattern, damping sources) and operating conditions [3].

Indeed, SHM methods make use of the vibration response of the structure to derive conclusions regarding its overall conditions (e.g. healthy or damaged). This is generally achieved by means of the acceleration signals gathered from a network of sensors positioned at specific points of the structure, and by subsequent processing to extract from acceleration signals information regarding the possible presence of damage (and also its localization and degree of severity).

In general, SHM can nowadays benefit from a suite of technologies from different fields that can be exploited by designers to develop effective systems at reduced costs. Indeed, recent advancements in the development of novel sensor devices, the advent of various communication and computation methodologies to support the so called Internet of Things (IoT), and the successful application of sophisticated machine learning tools represent three enabling factors for building successful SHM systems.

In the following, we provide a brief summary of the main contributions and key aspects regarding the progress of each of the analyzed technologies.

### A. Sensor Devices

Regarding the measurements of vibrations, accelerometers are commonly employed: they are in fact force-sensors coupled with a mass that applies a specific force proportional to the acceleration of the mass, according to the second Newton's law, when it is subject to vibration. Three types of transducers are mainly used to convert acceleration into proper electrical signals: piezoelectric, piezoresistive, and differential capacitive [1]. Piezoelectric accelerometers exploit the piezoelectric effect to measure changes occurring in specific materials undergoing mechanical stress. Piezoresistive sensors leverage the modifications induced in the electrical resistance of certain materials subject to mechanical deformation. Differential capacitive accelerometers measure changes in electrical capacitance to derive information about the displacement.

Regardless of the type of sensor, a common trend in many applications is the adoption of Micro-Electro-Mechanical-Systems (MEMS) as basic technology for SHM systems. Indeed, MEMS devices can be considered to be often on par with the performance of many macro-scale competitors, while they represent a lower cost and less invasive solution in many deployments.

### B. Processing and Communication Infrastructures

For what concerns communication infrastructures, SHM solutions have traditionally been based on wired networks.

However, the increased reliability of several wireless sensor networks (WSN) core technologies, integrated within a general Internet of Things communication protocols framework, have recently emerged as competitive alternatives thanks to their reduced installation costs and invasivity [2].

A standard IoT system usually consists of a three-tiered architecture with: *i*) a bottom layer, in charge of sensing and actuation through devices that collect data for transmission to upper layers and perform actuation according to specified policies/commands (potentially sent back from upper levels of the stack); *ii*) a middle layer, encompassing different types of communication networks (e.g. local area networks, cellular networks or, more generally, the internet) and devices (e.g. routers, gateways) enabled by different communication technologies, ranging from Bluetooth to WiFi, from LTE to 5G; *iii*) a top application layer devoted to data storage, processing and analysis usually hosted on cloud computing platforms.

### C. Machine Learning

Finally, once data is gathered from sensors (e.g. in the form of a time series) and is transmitted along the IoT communication pipeline, it is processed by means of machine learning techniques to perform inference and provide information (e.g. in the form of classification or regression) to downstream decision making. Deep learning represents the most recent forefront of this trend.

It should be remarked here the existence of an inherent tension between two different design strategies: on one hand, all data collected by a sensor can be transmitted remotely for its processing, on the other hand, it can be directly processed on board of sensor nodes. In the former case, data is sent for instance to edge devices, such as gateways, or to more powerful cloud computing hardware to which inference is delegated. In the latter case, machine learning models are used to perform predictions locally on peripheral devices. The above-mentioned tension is clearly the result of a trade-off among several variables, namely the energy spent for communication, the latency introduced by the processing/communication chain, and the accuracy achieved for solving a specific recognition task.

In this work, we introduce a vibration monitoring system purposed for SHM applications. The main goal of the proposed approach is to overcome the limitations of alternative solutions in terms of cost and flexibility (intended as ease of installation, maintenance, and operativity) through the design of a comprehensive system, which encompasses the following contributions:

- we design a novel Hall-effect based sensor to detect and measure vibrations; the resulting sensor is characterized by reduced costs and works at low frequency ranges (for instance it correctly detects peaks in the power spectrum of an earthquake waveform around 2.5Hz); the sensor is interfaced with off-the-shelf embedded platforms (i.e. a Raspberry PI) also characterized by reduced costs;
- we integrate the sensor node into an IoT framework that allows communicating readings to a cloud back-end system for visualization and processing purposes;

- we demonstrate the feasibility of performing complex machine learning inference directly on board of the adopted IoT hardware platform by implementing two different deep learning models (namely a convolutional and a recurrent neural network); non-trivial analysis of vibration signals by means of sophisticated deep learning models can be performed directly on board or offloaded to the cloud thanks to fully supported IoT functionalities;
- we provide an experimental assessment of the system under real-world working conditions in two use cases, namely the detection of a seismic event and an application for counting steps made by persons walking inside a public building; experimental results provide evidence of the versatility and flexibility of the proposed solution and its suitability as an alternative to other vibration monitoring systems.

The remainder of the article is organized as follows: in Section II we summarize the main state-of-the-art works in current scientific literature and frame their contributions with respect to our approach; in Section III we illustrate the proposed Hall-effect vibration sensor; in Section IV we describe the type and architecture of the two deep learning models used for performing inference; in Section V we introduce the setup adopted for the experimental validation of the proposed system and the related results obtained; in Section VI we conclude by recapitulating the main contributions and findings.

## II. RELATED WORK

The scientific literature about vibration monitoring in civil structures and infrastructures is particularly rich, being it a widely investigated subject of research.

A central assumption of SHM techniques is the possibility of monitoring the health status of a civil structure/infrastructure using vibration-based analysis. As a matter of fact, most SHM systems rely on the analysis of the so called-modal properties or vibration characteristics (i.e. natural frequencies, damping ratios, and mode shapes) which, in turn, depend on the physical features of the structure (i.e. mass, stiffness, and damping).

Since damaged structures change the physical properties of the structure, they result in modifications of the modal properties that can be used therefore as a proxy for damage detection [3]. Recent approaches have also introduced methods for detection and, at the same time, for the corresponding localization of possible damage within a structure, starting from the vibration data [4]. The main focus of our work is not the proposal of new methods for modal property analysis. Rather, we put forward a system-level design that could accommodate specific algorithmic approaches targeting vibration data analysis.

Importantly, many SHM methodologies nowadays make use of vibration data that can be extracted from MEMS sensors. These types of sensors, thanks to their increased precision, lower costs, and small size have progressively become a feasible alternative to traditional piezoelectric sensors [5], [2], [6]. In this article we describe a different approach for vibration sensing, based on a low-cost Hall-effect device which can be easily integrated into a wider IoT system.

Machine learning and, in particular, the deep learning paradigm, have increasingly gained popularity among scholars for vibration data processing [7]. For instance, in [8], the authors proposed a data-driven SHM method based on convolutional neural networks and fast Fourier transform to identify structural damage conditions from vibration data. An autoencoder architecture targeting nonlinear dimensionality reduction of input vibration signals has been recently introduced for the task of load identification in [9].

Another related line of research deals with the possibility of moving the execution of inference tasks (and also of learning tasks, in some cases) toward peripheral devices. Instead of relying on schemes entailing the transmission of whole data gathered from sensors to more powerful computing platforms (up to cloud systems), many works are currently investigating the opportunity to carry out specific machine learning tasks directly on board of sensor nodes or devices (such as gateways) located nearby in the network, according to *edge/fog computing* [10] or *tiny machine learning* paradigms [11], [12]. This potentially leads to benefits for security, privacy, and latency. The complex interplay between computation and communication also impacts energy consumption, which is a crucial factor to be taken into account in many IoT settings characterized by battery-operated devices[13].

Our contribution with respect to [8], [9] is represented by the implementation of two deep learning models on the embedded platform chosen as reference, hence showing the possibility of executing non-trivial machine learning tasks directly on the hardware platform in charge of collecting vibration data; given the integration with fully-fledged IoT protocols, the system can be reconfigured on the fly to also support cloud-enabled machine learning. We also benchmark the performances of the two implemented neural networks in a hypothetical smart building application (namely counting people's steps during walking).

Regarding the design of IoT systems for vibration monitoring, recently, Komarizadehal *et al.* have introduced a system based on an Arduino microcontroller equipped with five low-cost accelerometers [1]. The experimental assessment conducted by the authors with a comparison against two traditional piezoelectric sensors (as measured in terms of accuracy, resolution, and error) highlighted that the proposed platform can rival competitors, performing even better at low frequencies and low amplitude accelerations at a reduced (14x) cost [1]. In [14], an IoT sensing system for monitoring vibrations induced by construction sites is presented. The system, based on a Raspberry PI embedded platform and MEMS accelerometer is connected to the cloud via a USB dongle for 4G connectivity. Cloud-based frontend and backend complete the architecture implementing data storage and visualization (via MySQL database and web interface) and alarm detection based on predefined thresholds on the vibration signals.

Differently from our proposal, these works ([1], [14]) do not take into account any machine learning technique, while we also explore the implementation of machine learning models for vibration signal processing applications.

To conclude, the multifaceted field of SHM presents various challenges to be tackled by research toward the design of effective monitoring and diagnostics systems. Our proposal

attempts to bridge some gaps in the literature by contributing a low-cost solution based on a magnetic sensor device interfaced with a widely used commercially available hardware; the proposed solution supports machine learning inference on the device, as we demonstrate by embedding two deep learning models on it, while it also allows full integration with state-of-the-art IoT communication protocols to support cloud computing.

### III. THE PROPOSED HALL-EFFECT VIBRATION SENSOR

A Hall-effect sensor is a type of sensor in which the output voltage is directly proportional to the strength and direction of the magnetic field in which it is immersed. Its working principle is based on the Lorentz force induced by the presence of a magnetic field perpendicular to the direction of an electrical current applied to a thin strip of metal. As the result of the Lorentz force, a difference in electric potential (voltage) between the two sides of the strip can be measured which is proportional to the strength of the magnetic field [15]. As a special feature, a hall-effect sensor responds also to a static (non-changing) magnetic field differently from inductive sensors, which respond only to changes.

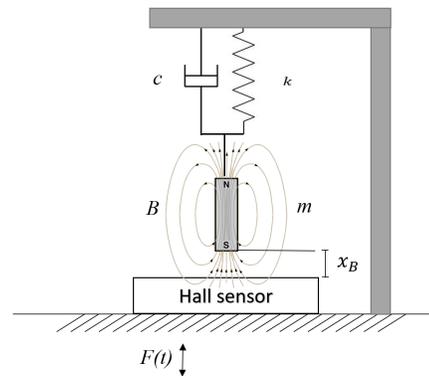


Fig. 1. The schematic representation of the proposed building vibration system consisting of a mass  $m$ , a spring  $k$ , a damping element  $c$ , and a magnetic field  $B$  positioned at a distance  $x_B$  from the hall-effect sensor

Hall sensors are traditionally used for proximity sensing, positioning, speed detection, and current sensing applications, or, combined with threshold detection, they act as a binary switch. In automotive, they are commonly used to time the speed of wheels and shafts, such as for internal combustion engine ignition timing, tachometers, and anti-lock braking systems.

In this work, we present a new low-cost IoT system, based on the Hall-effect, capable to measure building vibrations. In particular, we exploit the ability to produce an output proportional to the magnetic field of the Hall-sensor to dynamically monitor the displacement of an inertial static magnet with respect to the sensor which is rigidly anchored to the building.

Fig. 1 shows the main functioning principle of the proposed sensor. Notice that, for sake of simplicity the figure shows only the vertical axis but the same reasoning can be extended to other axes. The system can be led back to a classical damped harmonic oscillator consisting of a mass  $m$  suspended on a

spring characterized by a constant  $k$  together with a damping element with a viscous damping coefficient  $c$ . Moreover, in the proposed system, the suspended mass consists of a static magnet with a magnetic field  $B$  positioned at a distance  $x_B$  from a Hall-effect sensor. When the building is subject to vertical vibrations an external force  $F(t)$  is applied to the system which can be described, according to Newton's second law, by:

$$F(t) - kx - c \frac{dx}{dt} = m \frac{d^2x}{dt^2} \quad (1)$$

which can be rewritten into the form

$$\frac{d^2x}{dt^2} + 2\zeta\omega_0 \frac{dx}{dt} + \omega_0^2 x = \frac{F(t)}{m} \quad (2)$$

where  $\omega_0 = \sqrt{\frac{k}{m}}$  is called the undamped angular frequency of the oscillator and  $\zeta = \frac{c}{2\sqrt{mk}}$  is the damping ratio.

The value of the damping ratio  $\zeta$  critically determines the behavior of the system. A measuring system like the proposed one, must fall into the case called "underdamped" with  $\zeta < 1$ . Under this condition The system oscillates with the amplitude gradually decreasing to zero [16]. The angular frequency of the underdamped harmonic oscillator is given by

$$\omega_1 = \omega_0 \sqrt{1 - \zeta^2} \quad (3)$$

while the exponential decay of the harmonic oscillator is defined by

$$\lambda = \omega_0 \zeta \quad (4)$$

Considering a sinusoidal driving force defined as

$$F(t) = F_0 \sin(\omega t) \quad (5)$$

where  $F_0$  is the driving amplitude, and  $\omega$  is the driving frequency of the sinusoidal driving mechanism, equation 2 can be rewritten as:

$$\frac{d^2x}{dt^2} + 2\zeta\omega_0 \frac{dx}{dt} + \omega_0^2 x = \frac{1}{m} F_0 \sin \omega t \quad (6)$$

The general solution at the steady state that is independent of initial conditions and depends only on the driving amplitude  $F_0$ , driving frequency  $\omega$ , undamped angular frequency  $\omega_0$ , and the damping ratio  $\zeta$ . In particular, the steady-state solution is proportional to the driving force with an induced phase change  $\varphi$ :

$$x(t) = \frac{F_0}{mZ_m\omega} \sin \omega t + \varphi \quad (7)$$

where the value of the impedance of the system  $Z_m$  is defined as:

$$Z_m = \sqrt{(2\omega_0\zeta)^2 + \frac{1}{\omega^2}(\omega_0^2 - \omega^2)^2} \quad (8)$$

and the phase of the oscillation of the driving force  $\varphi$  is given by:

$$\varphi = \arctan\left(\frac{2\omega\omega_0\zeta}{\omega^2 - \omega_0^2}\right) + n\pi \quad (9)$$

Moreover, since the amplitude of the driving force  $F_0 = m\ddot{u}$  where  $\ddot{u}$  is the acceleration of the ground, the equation 7 can be rewritten as:

$$x(t) = \frac{\ddot{u}}{Z_m\omega} \sin \omega t + \varphi \quad (10)$$

From 10 results that the value of  $x(t)$  is proportional to the ground acceleration so measuring it entails measuring the ground acceleration. Furthermore, in the proposed system,  $x(t)$  describes  $x_B$  which is the distance between the hall sensor and the magnet and, since the measured voltage on the Hall sensor ( $V_{Hall}$ ) is proportional to the magnetic field  $B$ , which, in turn varies with the square of the distance, we can argue that measuring the voltage produced by the Hall sensor allows to measure the ground acceleration.

#### A. The Hardware Prototype

A prototype of the proposed system has been implemented by 3D printing a device capable of measuring building vibrations among the three orthogonal axes x, y, and z.

Fig. 2 shows the 3D solid model together with a picture of its realization. In particular, the central frame of the device sustains three orthogonal arms (yellow elements in the picture) that mount, at their ends, neodymium magnets. Each arm can only move along one degree of freedom as it is connected to the frame using a cylindrical bearing. The damping function is given by the bearing friction while the spring component has been implemented using magnetic repulsion with other neodymium magnets. Notice that, a calibrated mass has also been added to each arm since the very low mass of the magnets does not satisfy the design specifications. Finally, an analog linear Hall-effect sensor (Joy-it KY-024) has been placed perpendicularly and close to each suspended magnet to react at each displacement. The Hall-effect sensors are continuously sampled by means of a multichannel 16bit ADC (Az-Delivery ADS1115) which sends data to a Raspberry Pi model 3B+ through the I2C port.

Fig. 3 shows the circuit diagram together with a virtual representation of the electronic components. Since Hall-effect sensors produce an output also at a steady state (i.e. without any vibrations) the ADC has been connected in differential configuration so that it measures the difference between the sensor output and a reference value produced by means of the potentiometer R1. Tuning the potentiometer at a steady state allows calibrating the system to match the ADC specifications and to avoid ADC saturation.

#### B. The Management Software

The proposed hardware is managed by a python process installed as a daemon service in the Raspberry Pi which collects, elaborates, and sends data to a time series database.

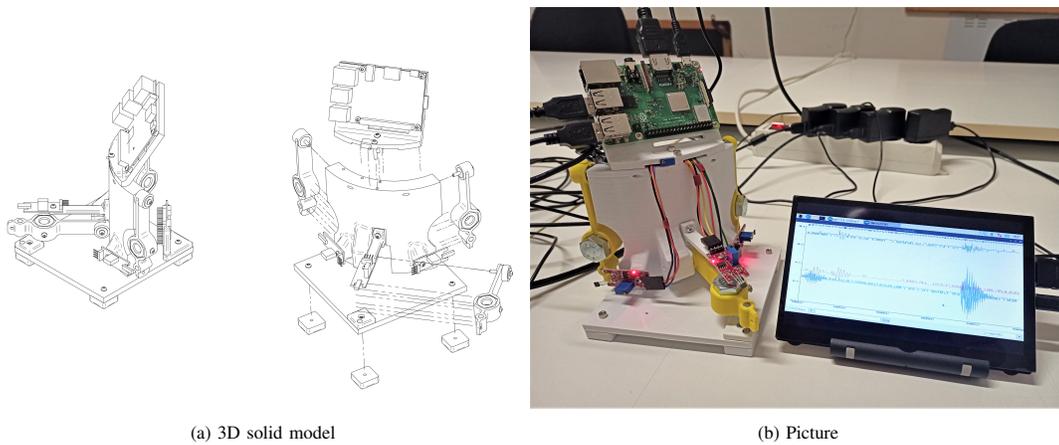


Fig. 2. A 3D solid model of the proposed system (a) together with a picture of its implementation (b)

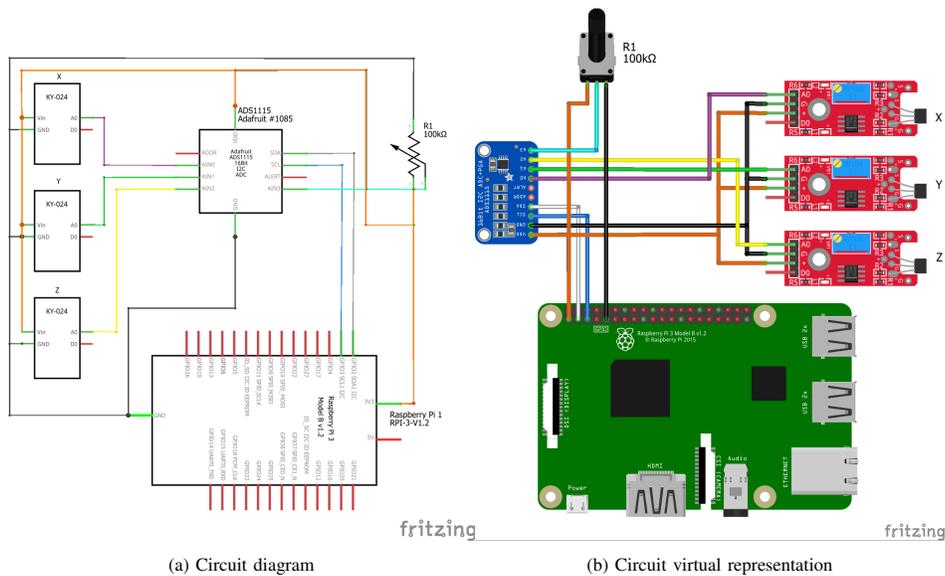


Fig. 3. A circuit diagram (a) and virtual representation (b) of the proposed sensor highlighting the single-board computer (raspberry Pi 3), the multichannel 16-bit ADC (ADS1115), and the three Hall-effect sensors (KY-024)

In particular, a first thread continuously samples the three ADC channels at a frequency of 50 Hz while a second thread elaborates and sends it to the remote server implementing a producer-consumer paradigm. In the current implementation, the consumer thread calculates the modulus of the three components of the signal and then computes the Root Mean Square (RMS) over a time window of 10 seconds. The RMS value is sent to the remote server while the entire buffer (10 seconds) containing the three signal components is stored only if the RMS exceeds a predefined threshold to avoid saving not significant data.

An autonomous monitoring system should guarantee the continuous collection of data avoiding as many as possible downtimes during which significant data could be lost. In a complex embedded system like the proposed one, several conditions can lead to the interruption of operation and the consequent loss of data such as, for example, a reboot of

the Raspberry Pi due to a glitch or a loss of the power supply, a crash on the python process, a problem on the I2C communication with the ADC module or a network down. To guarantee the best reliability of the monitor activity, the management software exploits several recovery strategies. First, installing the management process as a daemon service guarantees that the Linux OS automatically starts it at boot time or restarts the process in case of termination. Moreover, to recover the monitoring activity in case of a lack of data from the ADC, we activate a watchdog timer that reboots the platform if no data arrives for more than 10 seconds. Finally, the consumer thread has been designed to save data locally during network problems and, eventually, reestablish server communication when the network will work again.

#### IV. ADDING MACHINE LEARNING MODELS

To evaluate the suitability of the proposed system for the integration of machine learning algorithms, we tested two different supervised deep learning models, namely a Convolutional Neural Network (CNN) and a Long Short-Term Memory (LSTM) network aimed at recognizing a well-defined pattern, such as the vibrations induced by a walking person, within the time series collected by the system.

##### A. Deep-Learning Background

A CNN is typically composed of: *i*) a convolutional layer, *ii*) a pooling layer, *iii*) an activation function, *iv*) a fully connected layer, and *v*) a classification layer and it is traditionally used with very satisfying results in the field of image recognition and classification [17]. The convolutional layer takes as input a multi-channelled image, in  $n$ -dimensional format, and outputs a feature map. The pooling layer is in charge of keeping meaningful features while decreasing the size of the feature map. The activation layer allows us to learn other things, and its main role is to map the input to the output; the most well-known activation functions are: *Sigmoid*, *Tanh*, and *ReLU*, to cite a few. The fully connected layer acts as a classifier, and is normally placed at the end of the network, typically after the (last) pooling layer; the peculiarity of the fully connected layer is that each neuron is connected to all neurons of the predecessor layer. The final classification is however carried out by the classification layer, which is also responsible for error evaluation during training, i.e., for computing the difference between the predicted output and the actual one via the loss function. The most commonly adopted function is *Softmax*, which provides a prediction probability distribution.

It is worth noting that normally two-dimensional (2D) convolutional filters are used by CNN models to process 2D images; the opportunity to convert time series signal data to images is gaining a lot of momentum because it allows the application of computer vision techniques and performs classification tasks. In time series classification and pattern-recognition, this means that signals gathered from a triaxial sensor can be properly re-coded to images so that a “visual” analysis can be carried out to recognize, learn and classify patterns. Several re-coding techniques exist [18], [19], [20], and in this paper we leverage the deep learning approach proposed by [18], in which time series are converted to Gramian Angular Summation/Difference Fields (GASF/GADF) images. Such an approach envisages the representation of time series as a polar coordinate system that produce one image for the GASF and the other for the GADF. The main advantage of such an approach is that temporal and spatial relations are preserved.

Another deep learning approach used in time series classification is the LSTM model, which represents a recurrent neural network system. As detailed in [21], several versions of LSTM exist, but the most popular version is known as vanilla LSTM. A single LSTM unit is composed of a cell and three gates: input, output and forget gate. The cell is responsible for keeping track of values over time; the input gate is responsible for combining the current input, the output of the previous LSTM unit, and the value of the cell in the previous iteration

in order to decide whether to select the potential candidate values to be added. The forget gate considers current input, state on memory cells and output at previous time step in order to decide which information should be removed from previous cell states. The output gate computes the output (to be sent to the output block) by combining the output of that LSTM unit at previous time, the current input and cell value in the previous iteration. LSTM networks are able to capture temporal information from time series data; since CNN networks are able to automatically extract significant features, combining both networks to build a hybrid architecture can bring benefit in terms of performance and accuracy in time series pattern-recognition.

##### B. The Proposed Deep-Learning Models

Fig. 4 shows the architecture of the proposed LSTM (a) and CNN (b) models.

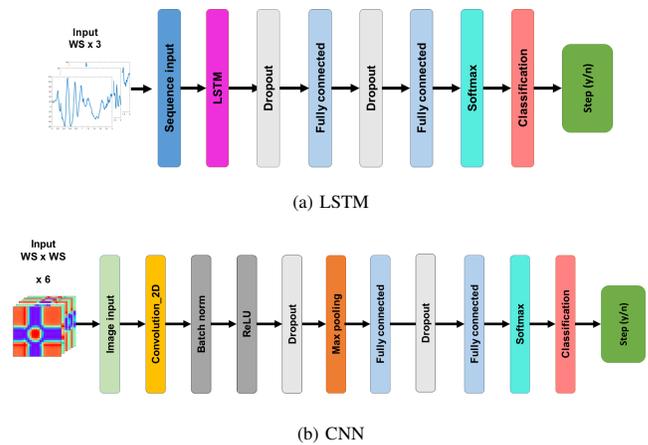


Fig. 4. The proposed deep learning models architecture i.e. (a) the LSTM network and (b) the CNN

Regarding the LSTM network, it takes in input the three signals gathered from the triaxial hall-sensor appropriately divided into windows of size equal to  $WS$ . These signals are arranged by a sequence input layer in a three-channel data chunk that is forwarded to an LSTM layer. The output of this layer is sent to two layers of fully connected neurons that act as a multi-layer Perceptron. Notice that, the two neurons layers are interspersed with two dropout layers, with a dropout probability 0.2, which helps prevent overfitting. The proposed network then terminates with a standard layer that computes the softmax function used by the classification layer to calculate the cross-entropy loss and to infer the presence or not of steps in the signal.

The proposed CNN model makes use of a the two-dimensional convolutional layer that acts on the six images obtained after encoding the three-time series into square images as described in Section IV-A. The images are then managed by an image input layer before going through the network structure. The convolutional layer is followed by a series of layers consisting of normalization, ReLU, dropout, and max-pooling layers. As in the previous network, the last part of the model is built by two neuron layers, a softmax, and a classification layer.

## V. EXPERIMENTAL SETUP AND RESULTS

In this section, we describe the experimental setup and we show the results obtained during two different sets of experiments, namely the building vibration monitoring and the machine learning-driven recognition of people's steps using the vibrations induced on the building floor.

### A. Experimental Setup

The Raspberry Pi model 3B+ was running the original Raspberry Pi OS (Raspbian) with an updated version of the Python3 interpreter. To manage the ADS115 16bit ADC we installed the *adafruit - circuitpython - ads1x15* library configured at the maximum data rate (860 samples per second). Thanks to the differential reading and the offset tuning circuit, the ADC gain was set to the maximum value which allows measuring in the  $\pm 0.256V$  range corresponding to a resolution of  $0.002V/bit$ .

In order to store, visualize, and process the collected data we built a time-series database by installing the InfluxDB open-source software on a desktop machine equipped with an Intel® Core i5 processor and 16GB of main memory running Ubuntu 22.04 LTS [22].

Concerning the machine learning experiments, the network models have been implemented, trained, and tested on the Matlab2022a® platform running on a Windows® desktop pc equipped with an Intel® Core i9 and 16GB of RAM. Moreover, once the models have been trained we converted them to C-language using the Matlab code-generation routine. The code was then compiled and executed on the Raspberry Pi to test its effectiveness and to measure the inference time. To provide a labeled dataset to the supervised classifier we manually annotated about 20 hours of collected traces with a binary label reporting the presence or not of vibrations induced by people's steps. Then the classification performances of the proposed classifiers have been measured using the following quantities:

$$Precision = \frac{TP}{TP + FP} \quad (11)$$

$$Recall = \frac{TP}{TP + FN} \quad (12)$$

$$F_1 score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \quad (13)$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (14)$$

where  $TP$  are the true positives,  $TN$  the true negatives,  $FP$  the false positives, and  $FN$  the false negatives.

### B. Measuring Building Vibrations

The prototype of the proposed monitoring system has been installed on a building named Collegio Raffaello located in the center of Urbino city - Italy. This building covers an area of about 2,600 square meters and it was built at the beginning of the eighteenth century by the will of Pope Clement XI. The

building floors are mainly made of wooden beams covered by ceramic tiles. Nowadays, the ground floor of the Collegio Raffaello holds some craft stores and a bar while on the first floor takes place the Physics Museum of the University of Urbino. Lastly, the second floor, where the proposed system has been positioned, currently houses the degree program in Applied Computer Science and the degree program in Foreign languages and cultures of the University of Urbino so that, during the lesson period, it is frequented by many students and teachers.

Fig. 5(a) shows the data extracted from the InfluxDB database which represents the RMS calculated over a 10 seconds time window during a week on May 2022. Traditionally the University lessons are held only from Monday to Tuesday and are organized in units of two hours where the first one starts at 9.00 a.m. and ends at 11.00 a.m.. Then a second unit begins which ends with the lunch break at 1.00 p.m. Finally, in the afternoon, there may be one or more units starting at 2.00 p.m. The lesson activity exactly matches the peaks reported in the figure which shows high RMS values during the daytime hours of the five days of activity while, in the last two days of the week, the building vibrations are practically absent. Furthermore, the vibrations recorded are maximum in central days (Wednesday and Thursday) as happens for the occupation of the building.

Since the system also sends to the remote server the single waveforms relating to the x, y, and z axes during vibrational events that exceed the set threshold, it is possible to extract these traces for subsequent processing. Fig. 5(b) shows the raw signals from the Hall-effect sensors expressed in volts which have been collected when a single person was walking on the floor near the system. As expected, the greatest activity is recorded on the vertical axis (z) due to the transfer of the weight of the person on the floor, but it is nevertheless interesting to note that even in the x and y directions, which are parallel to the plane, it is possible to record non-negligible vibrations due to the walking activity.

During continuous monitoring of the prototype a strong earthquake (Mw 5.5) struck Bosnia and Herzegovina region, causing the death of one person, the injury of at least two others, and forcing hundreds of people fleeing from their homes. Despite the distance that separates Urbino from the epicenter (about 450Km), the proposed monitoring system was able to record the waveforms related to the event for a duration of over 100 seconds. The raw signals are reported in Fig. 6 together with the ground acceleration collected by the official seismic station of the Italian National Seismic Network located at Monte Paganuccio (the closest station to Urbino). Despite the visible differences, it is interesting to note how the trend of the waveforms has several common points. For example, in both cases, the magnitude of the signal appears to be greater in the x and y axes, which are parallel to the floor, rather than in the vertical one. Moreover, both systems measure a vibration activity that lasts about 60 seconds. Clearly, a punctual comparison between the two systems, in addition to being out of the scope of this paper, is not possible for the fact that the system we proposed was located on the second floor of a building while the seismic station is directly supported on the ground so that the vibrations recorded by our system are, in effect, those induced on the building by the motion of the

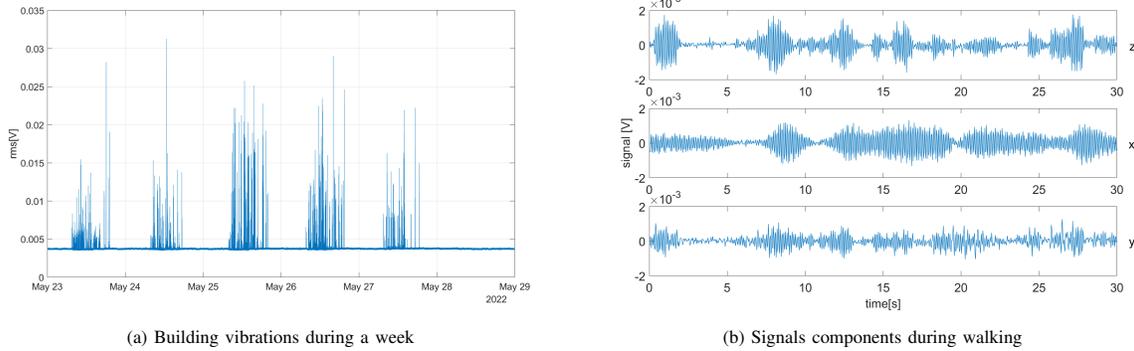


Fig. 5. Measured RMS of the building vibrations during an entire week (a) and triaxial components of the vibrations collected during single person walking

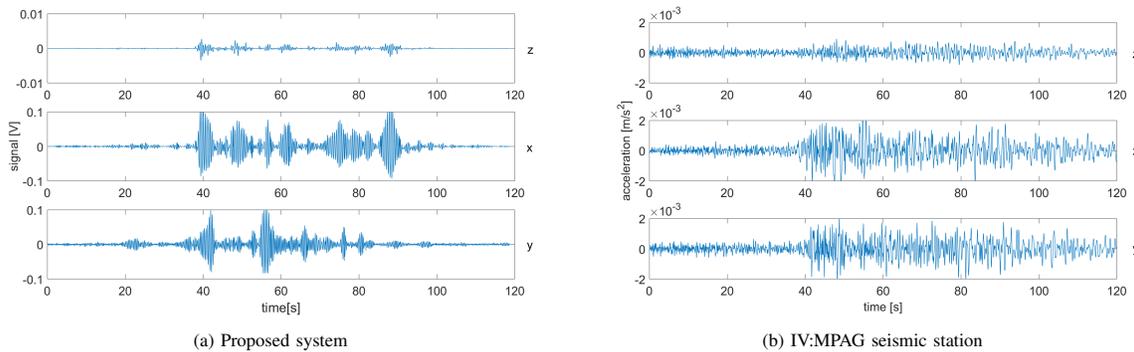


Fig. 6. Vibration waveforms measured during bosnia earthquake of April 22, 2022 (Mw 5.5) by the proposed system (a) and by the the IV:MPAG seismic station (b)

earthquake. Despite this, the comparison allows us to argue that the proposed system is sufficiently sensitive to low-intensity vibrations and therefore capable of monitoring the vibrations to which a building is subject.

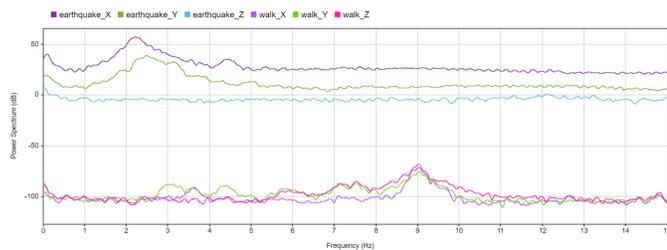


Fig. 7. Comparison of the spectral analysis of waveforms collected during an earthquake and single person walking

Finally, to evaluate the system response to the different patterns and kinds of vibrations, we report in Fig. 7 the power spectrum of the earthquake waveforms together with the spectrum of the vibration induced by a walking person. The first evident thing is that the two signals have very different predominant frequencies, for instance, the earthquake shows a main peak between 2 and 3 Hz (compatible with that detected by the seismic station) while the walking vibrations around 9 Hz.

### C. Detecting People's Steps

Starting from the labeled dataset, the two proposed models have been tested when varying the number of different components. The training set and test set have been split according to a holdout cross-validation methodology with 75% of the examples used for training and 25% kept for testing. Notice that, for a convolutional layer, this number represents the convolutional filters, for the LSTM layer the number of hidden units (i.e. the amount of information remembered between time steps), and for the fully connected layer the number of hidden neurons. Fig. 8 reports the performance measured after training and testing each configuration of the proposed models obtained when varying the number of internal components. As expected, both for LSTM and CNN, increasing the number of components increases the classification performances reaching a very high value (up to 96% of accuracy for LSTM). Each model was then ported on the Raspberry Pi to characterize it in terms of memory footprint and inference latency.

Table I reports the complete models characterization when varying the number of components. The increasing complexity of the models negatively reflects on the size and timing. Interestingly, the LSTM network outperforms the CNN both in terms of classification accuracy and memory and time requirements. For instance, the training time of CNN is 5× greater, the inference latency reaches about 3×, and the memory footprint is 7× greater with respect to the LSTM. Notice

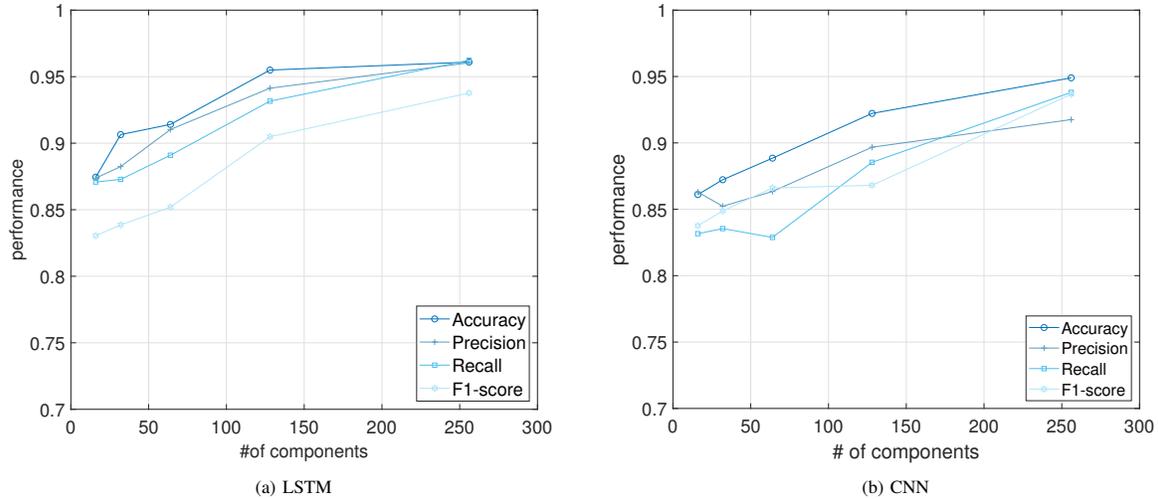


Fig. 8. Performance metrics of the LSTM (a) and CNN (b) networks when varying the network dimension

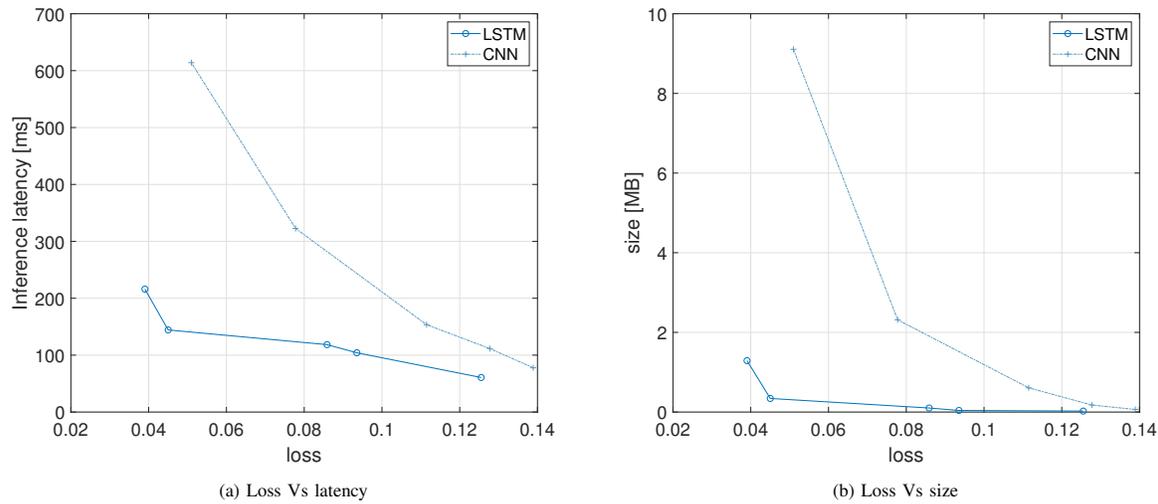


Fig. 9. Pareto curves of classification loss Vs inference latency (a) and model size (b)

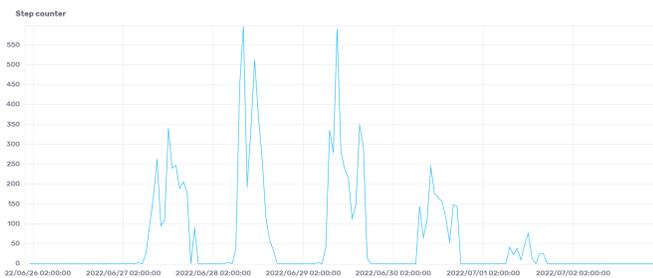


Fig. 10. The number of steps measured during a week on the collegio raffaello building

Pi as the inference latency would exceed the size of the monitoring window (500ms) thus making the consumer thread systematically slower than the producer with consequent data loss.

In Fig. 9, the Pareto curves obtained by plotting the classification loss Vs the inference latency (a) and model size (b) are reported. It is evident that the high classification performance combined with the low inference latency and the very low memory occupation make the LSTM network the best compromise for the detection of people's steps starting from the vibrations induced on the building using the proposed system.

that, for real-time requirements the last CNN configuration (i.e. # of components = 256) cannot be used on Raspberry

According to these results, we permanently installed the trained LSTM on our system prototype with the scope of measuring the walking activity on the building floor.

TABLE I. CNN AND LSTM ACCURACY, TRAINING TIME, LATENCY, AND MEMORY FOOTPRINT AS FUNCTION OF NUMBER OF COMPONENTS

# of components	Accuracy		Training T [s]		Latency T [ms]		Size [kb]	
	LSTM	CNN	LSTM	CNN	LSTM	CNN	LSTM	CNN
16	0.874	0.861	950	4534	60.7	77.6	25	68
32	0.906	0.872	818	4398	103.2	111.5	42	184
64	0.914	0.889	844	4449	114.8	153.4	106	638
128	0.955	0.922	868	4623	144.3	322.4	358	2429
256	0.961	0.949	921	5127	217.2	614.6	1353	9550

Fig. 10 shows the number of steps measured during the last week of June 2022. Notice that, this week there was no teaching activity so the building was frequented only by teachers. In particular, the plot has been obtained using the influxdb server querying language to sum the number of counted steps over a time window of one hour. The maximum value of about 550 steps/hour was found on Tuesday and Thursday which are the days of greatest attendance by teachers while the total amount of recorded steps was about 6,500 steps.

## VI. CONCLUSIONS

Vibration monitoring systems are traditionally used to assess the vibration impact of natural and human activities on buildings, to evaluate the health condition of structures. However, classical vibration monitoring systems are associated with limitations such as expensive devices, difficult installation, complex operation, etc. In this paper, we presented a low-cost Internet of Things monitoring system based on Hall-effect sensors encompassing integration with deep-learning machine models. The results obtained in real-life experiments (namely, the detection of seismic events and the count of steps made by people during walking in a public building) demonstrate that the proposed system can effectively gather vibration data in a built environment and operate in a standard fashion by transmitting all data remotely toward cloud-base backend or, conversely, perform intelligent pattern recognition tasks directly on the end device with accurate performance levels and reduced latency. We may therefore conclude that the proposed solution could represent a versatile and promising alternative to traditional vibration monitoring systems.

## REFERENCES

[1] S. Komarizadehasl, B. Mobaraki, H. Ma, J.-A. Lozano-Galant, and J. Turmo, "Development of a low-cost system for the accurate measurement of structural vibrations," *Sensors*, vol. 21, no. 18, p. 6191, 2021.

[2] F. Di Nuzzo, D. Brunelli, T. Polonelli, and L. Benini, "Structural health monitoring system with narrowband iot and mems sensors," *IEEE Sensors Journal*, vol. 21, no. 14, pp. 16371–16380, 2021.

[3] R. Astroza, H. Ebrahimian, J. P. Conte, J. I. Restrepo, and T. C. Hutchinson, "Statistical analysis of the modal properties of a seismically-damaged five-story rc building identified using ambient vibration data," *Journal of Building Engineering*, vol. 52, p. 104411, 2022.

[4] Y. Liao, A. S. Kiremidjian, R. Rajagopal, and C.-H. Loh, "Structural damage detection and localization with unknown postdamage feature distribution using sequential change-point detection method," *Journal of Aerospace Engineering*, vol. 32, no. 2, p. 04018149, 2019.

[5] S. M. Khan, M. U. Hanif, A. Khan, M. U. Hassan, A. Javanmardi, and A. Ahmad, "Damage assessment of reinforced concrete beams using cost-effective mems accelerometers," *Structures*, vol. 41, pp. 602–618, 2022. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S2352012422003629>

[6] F. Zonzini, M. M. Malatesta, D. Bogomolov, N. Testoni, A. Marzani, and L. De Marchi, "Vibration-based shm with upscalable and low-cost sensor networks," *IEEE Transactions on Instrumentation and Measurement*, vol. 69, no. 10, pp. 7990–7998, 2020.

[7] M. Flah, I. Nunez, W. Ben Chaabene, and M. L. Nehdi, "Machine learning algorithms in civil structural health monitoring: a systematic review," *Archives of computational methods in engineering*, vol. 28, no. 4, pp. 2621–2643, 2021.

[8] Y. He, H. Chen, D. Liu, and L. Zhang, "A framework of structural damage detection for civil structures using fast fourier transform and deep convolutional neural networks," *Applied Sciences*, vol. 11, no. 19, p. 9345, 2021.

[9] L. Rosafalco, A. Manzoni, S. Mariani, and A. Corigliano, "An autoencoder-based deep learning approach for load identification in structural dynamics," *Sensors*, vol. 21, no. 12, p. 4207, 2021.

[10] O. Gómez-Carmona, D. Casado-Mansilla, F. A. Kraemer, D. López-de Ipiña, and J. García-Zubia, "Exploring the computational cost of machine learning at the edge for human-centric internet of things," *Future Generation Computer Systems*, vol. 112, pp. 670–683, 2020.

[11] E. Lattanzi, M. Donati, and V. Freschi, "Exploring artificial neural networks efficiency in tiny wearable devices for human activity recognition," *Sensors*, vol. 22, no. 7, p. 2637, 2022.

[12] R. David, J. Duke, A. Jain, V. Janapa Reddi, N. Jeffries, J. Li, N. Kreeger, I. Nappier, M. Natraj, T. Wang *et al.*, "Tensorflow lite micro: Embedded machine learning for tinyml systems," *Proceedings of Machine Learning and Systems*, vol. 3, pp. 800–811, 2021.

[13] A. Bogliolo, E. Lattanzi, and V. Freschi, "Idleness as a resource in energy-neutral wsns," in *Proceedings of the 1st International Workshop on Energy Neutral Sensing Systems*, 2013, pp. 1–6.

[14] Q. Meng and S. Zhu, "Developing iot sensing system for construction-induced vibration monitoring and impact assessment," *Sensors*, vol. 20, no. 21, p. 6120, 2020.

[15] E. Ramsden, *Hall-effect sensors: theory and application*. Elsevier, 2011.

[16] J. P. Den Hartog, *Mechanical vibrations*. Courier Corporation, 1985.

[17] L. Alzubaidi, J. Zhang, A. J. Humaidi, A. Al-Dujaili, Y. Duan, O. Al-Shamma, J. Santamaría, M. A. Fadel, M. Al-Amidie, and L. Farhan, "Review of deep learning: Concepts, cnn architectures, challenges, applications, future directions," *Journal of big Data*, vol. 8, no. 1, pp. 1–74, 2021.

[18] Z. Wang and T. Oates, "Imaging time-series to improve classification and imputation," in *IJCAI International Joint Conference on Artificial Intelligence*, vol. 2015-January, 2015, pp. 3939–3945.

[19] G. Baldini, G. Steri, R. Giuliani, and C. Gentile, "Imaging time series for internet of things radio frequency fingerprinting," in *2017 International Carnahan Conference on Security Technology (ICCSST)*. IEEE, 2017, pp. 1–6.

[20] Z. Qin, Y. Zhang, S. Meng, Z. Qin, and K.-K. R. Choo, "Imaging and fusing time series for wearable sensor-based human activity recognition," *Information Fusion*, vol. 53, pp. 80–87, 2020.

[21] G. Van Houdt, C. Mosquera, and G. Nápoles, "A review on the long short-term memory model," *Artificial Intelligence Review*, vol. 53, no. 8, pp. 5929–5955, 2020.

[22] InfluxData. (2022) InfluxDB: Open Source Time Series Database. [Online]. Available: <https://www.influxdata.com/>

# Sequence Recommendation based on Deep Learning

Gulsim Rysbayeva<sup>1</sup>, Jingwei Zhang<sup>2</sup>

School of Computer Science and Information Security, Guilin University of Electronic Technology, No.1, Jinji Rd.,  
Guilin 541004, Guangxi<sup>1</sup>  
Guangxi Key Laboratory of Trusted Software, Guilin University of Electronic Technology,  
Guilin 541004, China<sup>2</sup>

**Abstract**—Sequence recommendation systems have become increasingly popular in various fields such as movies and social media. These systems aim to predict a user's preferences and interests based on their past behavior and provide them with personalized recommendations. Deep learning, particularly Recurrent Neural Networks (RNNs), have emerged as a powerful tool for sequence recommendation. In this research, we explore the effectiveness of RNNs in movie and Instagram recommendation systems. We investigate and compare the performance of different types of RNNs, such as Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU), in recommending movies and Instagram posts to users based on their browsing history. Additionally, we study the impact of incorporating additional information such as user's demographics and Instagram hashtags on the performance of the recommendation system. We also evaluate the performance of RNN-based movie and Instagram recommendation systems in comparison to traditional approaches, such as collaborative filtering and content-based filtering, in terms of accuracy and personalization. The findings of this research provide insights into the effectiveness of RNNs in movie and Instagram recommendation systems and contribute to the development of more accurate and personalized recommendations for users

**Keywords**—Long short-term memory (LSTM) and gated recurrent unit (GRU); RNN; deep learning; recommendation systems

## I. INTRODUCTION

Sequence recommendation systems have become increasingly popular in various fields, such as movies and social media. These systems aim to predict a user's preferences and interests based on their past behavior and provide them with personalized recommendations. One of the most promising approaches for sequence recommendation is deep learning, which has been shown to be effective in capturing the complex patterns and dependencies in sequential data.

Sequence recommendation systems have become an important area of research in recent years due to the growing amount of sequential data available in various fields such as music, videos, e-commerce, and social media. These systems aim to predict a user's preferences and interests based on their past behavior and provide them with personalized recommendations. Traditional approaches to sequence recommendation, such as collaborative filtering and content-based filtering, have been shown to be effective to some extent. However, they are limited in their ability to capture the complex patterns and dependencies in sequential data.

Deep learning, particularly Recurrent Neural Networks (RNNs), have emerged as a powerful tool for sequence recommendation. RNNs are particularly suitable for this task as they are able to model the temporal dependencies in the user's browsing history. Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) are two popular types of RNNs that have been widely used in sequence recommendation systems. LSTMs have been shown to be effective in capturing long-term dependencies in sequential data, while GRUs have been shown to be more computationally efficient.

Recent research has explored the application of RNNs in movie and Instagram recommendation systems. For example, researchers have used RNNs to model user's browsing history and predict their preferences for movies. Other researchers have used RNNs to model user's interactions on Instagram and recommend posts based on their interests. However, there is still a need to further investigate the effectiveness of RNNs in these systems and explore how incorporating additional information such as user's demographics, and Instagram hashtags can improve the performance of the recommendation system.

This research aims to fill this gap by exploring the effectiveness of RNNs in movie and Instagram recommendation systems and investigating the impact of incorporating additional information on the performance of the system. The findings of this research will contribute to the development of more accurate and personalized recommendations for users.

The vast amount of sequential data available in these domains, such as users browsing history, provides an opportunity for recommendation systems to offer personalized recommendations to users. However, despite the growing interest in using RNNs for sequence recommendation, there is still a lack of understanding of how well they perform in these specific domains and how to effectively incorporate additional information to improve the performance of the system. In particular, traditional approaches such as collaborative filtering and content-based filtering have been shown to be effective to some extent, but they are limited in their ability to capture the complex patterns and dependencies in sequential data, while RNNs, specifically LSTMs and GRUs, have shown to be promising in modeling temporal dependencies. This research aims to explore the effectiveness of RNNs in movie and Instagram recommendation systems by comparing the performance of different types of RNNs, and investigate the impact of incorporating additional information such as

user's demographics and Instagram hashtags on the performance of the recommendation system. Additionally, the research aims to compare the performance of RNNs-based movie and Instagram recommendation systems with traditional approaches in terms of accuracy and personalization.

In this research paper, we will focus on exploring the effectiveness of Recurrent Neural Networks (RNNs) in movie and Instagram recommendation systems. Movie recommendation is a challenging task because of the large number of movies available, the diversity of genres, and the dynamic nature of user preferences. Similarly, Instagram recommendation is a challenging task because of the large number of posts and users available, the diversity of content and the dynamic nature of user preferences. RNNs are particularly suitable for these tasks as they are able to model the temporal dependencies in the user's browsing history. We will investigate and compare the performance of different types of RNNs, such as Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU), in recommending movies and Instagram posts to users. Additionally, we will also explore the impact of incorporating additional information such as user's demographics, and Instagram hashtags on the performance of the recommendation system.

The research in this paper will provide insights into the effectiveness of RNNs in movie and Instagram recommendation systems and contribute to the development of more accurate and personalized recommendations for users. The objectives of this research are as follows:

- To investigate and compare the performance of different types of Recurrent Neural Networks (RNNs), such as Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU), in recommending movies and Instagram posts to users based on their browsing history.
- To study the impact of incorporating additional information such as user's demographics and Instagram hashtags on the performance of the recommendation system.
- To provide insights into the effectiveness of RNNs in movie and Instagram recommendation systems and contribute to the development of more accurate and personalized recommendations for users.
- To identify any limitation that the proposed RNN-based recommendation systems may have, and propose potential solutions to overcome them.

The paper is divided into five sections, which are organized as follows. The first section is the introduction, which provides an overview of the research topic and the research questions that the paper aims to address. In this section, the paper also provides background information on the topic and explains the significance of the research. The second section is the related work, which reviews the existing literature on the research topic. This section provides an overview of the key findings and ideas from previous research and explains how these findings relate to the research

questions addressed in the paper. The third section is the methodology, which explains the research design and methods used to collect and analyze the data. This section provides detailed information on the research participants, data collection procedures, and data analysis techniques. The fourth section is the results, which presents the findings of the research. This section provides an overview of the key findings, including any statistical analyses that were performed, and discusses the implications of the findings for the research questions. The fifth and final section is the conclusions, which summarizes the key findings of the research and draws conclusions about the research questions. This section also discusses the limitations of the research and suggests directions for future research.

## II. RELATED WORK

Algorithms help recommender systems make personalised suggestions. These systems now use artificial intelligence-based machine learning. The literature mentions various machine learning algorithms for recommender systems, making it hard to choose one. Recommender system researchers know nothing about algorithm usage. Developing recommender systems with machine learning algorithms often encounters issues. This paper [1] analysed machine learning research on recommender systems, identifying gaps and suggesting future research. This study aims to help new researchers appropriately situate their research by identifying trends in the use or study of machine learning algorithms in recommender systems and unresolved difficulties. Current categories of recommender systems are identified, selected machine learning techniques are characterised, big data technologies are discussed, different types of machine learning algorithms and their best domains are identified, and both major and alternative performance metrics are investigated.

Online stores increasingly use recommendation systems. Current recommendation algorithms are good at optimising a single task (such as click-through rate prediction) based on users' historical click sequences, but they don't model users' multiple behaviours or jointly optimise multiple objectives (such as both CTR and Conversion rate), which are crucial for e-commerce sites. This research suggested that multi-task learning and user interests based on a wide range of behaviours are needed to make meaningful progress towards multiple goals at once. This study [2] introduced Deep Multifaceted Transformers (DMT), a system that uses many Transformers to express multiple user behaviour sequences. "Multi-gate Mixture-of-Experts" does this.

Sequential recommendation research is growing. This study [3] recommends media and online products to customers based on their past behaviour and interests. Recent machine learning algorithms for sequential recommendation use deep learning and Transformers. In view of the surprising competitive performance of basic nearest-neighbor algorithms for session-based recommendation, the author examined nearest-neighbor techniques for sequential recommendation challenges. For two of four datasets, nearest-neighbor approaches outperformed the Transformer-based BERT4REC algorithm. Deep learning surpasses the simpler approaches for

the two bigger datasets, supporting the idea that neural methods improve with data.

A common recommendation situation made attainable by the internet has greatly enhanced user data collection. Authors can predict the user's next action by evaluating latent heterogeneous collaborative signals and sequential patterns. Sequential recommendation approaches and heterogeneous information network-based methods worsen the common data sparsity problem [4]. This innovative Sequence-aware Heterogeneous graph neural Collaborative Filtering (SHCF) model solves these problems by considering both high-order heterogeneous collaborative signals and sequential information.

Recent scholars have focused on sequential recommender systems, a nascent field (SRSs). SRSs model sequential user behaviours, user-thing interactions, and the temporal evolution of user preferences and item popularity, unlike collaborative filtering and content-based filtering. SRSs characterise user settings to provide more accurate, tailored, and dynamic recommendations. Researchers are becoming interested in sequential recommender systems (SRSs). SRSs represent and interpret sequential user behaviours, object-user interactions, and the time evolution of user preferences and item popularity, unlike collaborative filtering and content-based filtering. SRSs leverage the above attributes to better reflect user settings, interest and goals, and item consumption behaviour to deliver more accurate, tailored, and adaptable suggestions. This study [5] explained SRSs. Author defined SRSs, analysed and classified the most significant difficulties in this field, and reviewed SRS research progress, including present and past successes. Lastly, the author proposed intriguing study directions for this emerging discipline trend in such items consumption and your goals. This research introduces SRSs thoroughly. Author defined SRSs, evaluated and categorised the most urgent issues facing this area of study, and then examined SRS research progress, including the most recent and typical triumphs in this subject. Finally, the author suggests prospective research areas in this dynamic field.

DL-based sequential suggestion models have surpassed Markov chain-based and factorization-based methods in recent years. Yet, a reliable DL model for sequential recommendation has been shockingly neglected. This work [6] addresses deep learning-based sequential recommender systems for these difficulties. This study summarised the main factors that affect the performance of DL-based models and performed corresponding evaluations to showcase and demonstrate their effects. It also illustrated sequential recommendation, proposed a categorization of existing algorithms into three types of behavioural sequences, and demonstrated the effects of these factors. Finally, the author examined the sector's opportunities and challenges.

Sequential recommendation systems leverage user behaviour to anticipate preferences. Due to data paucity, this [7] strategy may not work as well in practise. This study introduced counterfactual data augmentation to help sequential recommendation algorithms overcome incomplete training data. The sampler and anchor models comprise this architecture. Anchor models train the final recommendation

list using both observed and produced sequences, whereas sampler models develop new user behaviour sequences based on the observed ones. This sampling method asks what a customer might buy if her recent purchases were different. Author used two learning-based methods to create the sampler model and increase output sequence quality while training the anchor model. To flesh out the image, the author theoretically evaluated how the generated sequences affected the anchor model, finding a balance between information and noise. This method was evaluated on nine real-world datasets, proving its usefulness and generalizability.

Recommender systems are increasingly used. Amazon and eBay offer millions of products for customers. Buyers can choose from a huge selection of products online. This increased level of personalization requires buyers to sift through a lot of company data. Recommendation systems can help manage this information overload. Conventional recommender systems use prior activities and profiles to provide recommendations. Deep learning methods have performed well in many situations. Yet, deep learning applications in recommendation systems have not been extensively studied. This tutorial's first segment introduced recommender systems and deep learning. In the following chapters, this study [8] reviewed and critiqued many state-of-the-art deep recommendation systems.

Due to the volume of products and consumers' changeable tastes, fashion-focused e-commerce businesses need recommendation engines. Since most users browse anonymously, historical preference data is rarely available, therefore suggestions can only be based on current session data. Dressipi rented 1.1 million fashion e-commerce sessions over 18 months for the 2022 ACM RecSys competition. Predicting a consumer's basket by session's conclusion was the goal [9]. All sessions are private and never save user data to replicate a real-world production setting. This article describes author's solution. Self-supervised learning inspired the Transformer design with dual learning goals to boost generalisation.

Online service domains are increasingly using sequential suggestion. It models consumer tastes based on past behaviour to predict spending. Real-world systems may capture massive amounts of user activity. This vast dataset can reveal consumer preferences. Hence, past programmes have largely recommended future behaviour based on observed patterns. Due to past experiences influencing current choices, sequential data may not be fully utilised. Real-world systems need speed, thus it's no longer practical to watch user behaviour before making conclusions. The Dynamic Memory-based Attention Network, a revolutionary long-sequence-based recommendation model, addresses this issue (DMAN). It breaks down the user's prior behaviour, trains the model, and saves memory blocks to preserve their preferences. DMAN dynamically splits each user's long-term interests into memory blocks to minimise auxiliary reconstruction loss and maximise memory integrity. Dynamic memory prepares the user's immediate and delayed preferences for a single set of suggestions. Empirical results reveal that this work [10] outperforms state-of-the-art sequential models on four benchmark datasets in capturing long-term dependencies.

User modelling using sequential suggestion reverses the user's interaction history to determine preferences. Such techniques require solid datasets with real sequential information for evaluation. This study [11] examines the timestamps of several major datasets and finds no meaningful sequential order. The datasets show that multiple user interactions occur at the same time. The dialogue between the parties is only semi-sequential. A leading sequential recommender performs similarly when encounters are randomly reordered. This page discusses Cinema Lens. Authors have noticed that sequential recommenders require new datasets with better sorting.

RNNs model sequences well. They can be enlarged, hold different types of information, and account for time. They're great at making sequential recommendations. This study [12] adds Recommender System considerations to Recurrent Neural Networks. The recommendation recipient's clear image is one such aspect. Author demonstrated how a novel Gated Recurrent Unit may efficiently represent consumers and their consumption habits to make personalised buy recommendations. These upgrades outperformed state-of-the-art recommender algorithms and a baseline Recurrent Neural Network in offline tests on two real-world datasets.

Sequential recommenders struggle to predict user behaviour as they move from recommendation to recommendation. "Cold-start" consumers with few real-world encounters are the problem. The difficulty of learning sequential patterns over consumers with few encounters will diminish the predictive power of sequential recommendation algorithms. MetaTL, a novel framework that models user transitions using meta-learning, improves sequential suggestion for novice users [13]. In particular, MetaTL 1 formulates sequential recommendation for cold-start users as a few-shot learning problem, extracts dynamic transition patterns among users with a translation-based architecture, and uses meta transitional learning to enable fast learning for cold-start users with limited interactions, resulting in accurate inference of sequential interactions.

In e-commerce, the ability to recommend products based on previous purchases can boost income. Most recommender systems ignore reviews, which include a wealth of information about the user's tastes. This study [14] compares 10 RNN architectures to make user-evaluated suggestions. We have explored and built multi-stacked bi-directional Gated Recurrent Units (GRUs), Long Short-Term Memory (LSTMs), and other RNN architectures.

Information overload necessitates recommendation systems. It may reduce research time by offering recommendations and forecasts based on user behaviour. More deep learning research uses deep neural networks. This study [15] introduced recommendation systems and deep learning algorithms. Each recommendation system is detailed and linked to its data sets. This paper's 2014–2017 citations provide a current overview.

Due to the sheer volume of online information, users may have trouble filtering, prioritising, and communicating pertinent material. Recommender systems employ real-time data to make customised recommendations. This study [16]

presents a smart clothing recommendation system to address the increased demand for customised outfits. Transfer learning extracts cosine similarity data from product photos to customise recommendations. In an online clothes store, the author exploited this. Displaying photos that are 80% or more like the user's product solves the personalised suggestion problem.

User modelling powers online recommendation systems. CF approaches are widely used to model customers' long-term preferences. Recurrent neural networks (RNN) have improved short-term customer choice prediction. Mixing long- and short-term models strengthens suggestions. Previous methods neglected dynamically merging these two user modelling paradigms. Traditional RNN structures like Long Short-Term Memory (LSTM) can handle linguistic and visual input, but they need to be improved to account for user behaviour. This study [17] proposes a time-aware and content-aware controller to improve RNN architecture by incorporating contextual information during state transitions.

Matching, or detecting if a document is relevant to a user's search query or interest level, is a major difficulty in search and recommendation. Machine learning, or "learning to match," has been applied to the issue using input representations and labelled data. Deep learning algorithms have been investigated to improve search and recommendation matching performance. Deep learning for matching has become the search and recommendation standard due to an unprecedented data influx, powerful computational resources, and cutting-edge deep learning techniques. Deep learning must learn representations and generalise data-matching patterns for this investigation [18].

Internet information overload? Recommendation systems support. With their extensive use in online applications and ability to mitigate over-choice issues, recommender systems are invaluable. Deep learning has performed well in recent years and can learn feature representations from scratch, attracting academics from many domains. Deep learning has been effectively used to information retrieval and recommender system studies, expanding its influence. Deep learning in recommender systems is growing. This work [19] aimed to review recent deep learning studies on recommender systems. The author assessed current methodologies and provided a taxonomy of deep learning-based recommendation models. Lastly, author added additional insights into this amazing development.

In recent years, deep learning has outperformed Markov chain and factorization-based sequential recommendation models. However, DL-based approaches have substantial limitations, such as inadequate user representation modelling and a failure to discriminate between user behaviour and object interactions. This research [20] addressed these concerns with sequential recommender systems built on DL. Examples of sequential recommendations, a categorization of existing algorithms based on three behavioural sequences, factors impacting DL-based model performance, and matching experiments are shown. Finally, author mapped out the future paths and problems of this discipline.

In this study [21], author describes five previously untested ways for strengthening deep learning-based top-n suggestions. A "Collaborative Memory Network" uses the latent component model and neighborhood-based algorithms to collaborative filter using implicit input, inspired by the Memory Network. Next, the author presented Neural Semantic Personalized Ranking, a probabilistic generative modelling approach that combines deep neural networks with paired rankings to solve this problem. Finally, author suggested adding a context-driven attention technique to Attentive Contextual Denoising Auto encoder for unstructured user and product data. The author created the context-aware Neural Citation Network using a flexible encoder-decoder architecture. This system uses an effective max time delay neural network encoder, attention mechanism, and author networks. To finish up, author proposed a general framework for natural language processing-based user preference inference utilising transfer learning for movie choices during conversation.

As internet data grows, recommender systems help manage information overload. Recommender systems are prevalent in online applications and can solve many difficulties caused by too many choices. Deep learning's amazing performance and the enticing property of learning feature representations from scratch have gained a lot of attention from a broad array of academics in disciplines including computer vision and natural language processing in recent years. Lately, deep learning has also had an effect on the research of recommendation and information retrieval systems, indicating its utility in this arena. The use of deep learning to recommender systems is a rapidly emerging subject at the moment. This study's [22] main purpose was to give readers with a comprehensive evaluation of the current findings from works on deep learning recommender systems. This publication provided an extensive overview of deep learning-based recommendation models and a taxonomy for grouping them.

As the number of individuals with Internet access grows, as does the need for personalised experiences and the speed at which people's online habits change, recommender systems have become a viable way to sort through massive volumes of data to find the most relevant results. Modern recommender systems produce useful recommendations but have accuracy, scalability, and cold-start concerns. Deep learning and other modern machine learning techniques, used for a wide range of complicated tasks, have been applied to recommender systems to improve their recommendations. For beginners, this study [23] analysed deep learning-based recommendation systems in detail. Author evaluated literature on deep learning models for recommender systems, obstacles, recommendation domain knowledge, and function. The author quantified all relevant studies and analysed these findings and ideas for additional investigation.

YouTube is a leading recommendation system. This study [24] provided a system overview and highlighted deep learning's efficiency gains. The research describes two unique deep candidates generating and ranking algorithms, and it does so in accordance with the standard two-stage paradigm for the retrieval of data. Author shared the expertise that

author earned via the process of planning, creating, and maintaining a big recommendation system that has a discernible impact on end users.

The area of machine learning had substantial success in several domains [25]–[27] and as a whole it has been dramatically changed by the emergence of deep learning. To be sure, its impact on the area of recommender systems wasn't immediately apparent. Yet, it was significant. This paper [28] discussed Netflix's issues employing deep learning for recommender systems and the lessons learned. Author first listed all Netflix suggestion-related occupations. Author found that different model designs work in different settings. Despite the fact that many deep-learning models may be considered as extensions of preexisting (basic) recommendation algorithms, author did not originally find large increases in performance over well-tuned alternatives that did not apply deep learning. Many deep-learning models are extensions of current approaches, although not all. Deep learning models were ineffective until the author added a significant number of characteristics from several heterogeneous data sources. Deep learning can worsen offline-online metrics misalignment, according to authors. Deep learning improved this recommendation by old and current metrics. This happened after author fixed deep learning issues.

### III. PROPOSED METHODOLOGY

The proposed model for this study is a Recurrent Neural Network (RNN) based movie and Instagram recommendation system. The model will consist of three main components:

- An encoder that processes the user's browsing history and extracts relevant features. This component will likely use a type of RNN, such as LSTM or GRU, to capture the temporal dependencies in the user's browsing history.
- An attention mechanism that assigns a weight to each feature based on its relevance to the user's preferences. This component will help the model to focus on the most important features of the user's browsing history.
- A decoder that generates recommendations based on the encoded features and attention weights. This component will likely use a type of RNN, such as LSTM or GRU, to generate the recommendations.

RNNs are powerful tools for processing sequential data such as time series, natural language, and user behavior. They can capture the temporal dependencies and patterns within the sequences and use that information to make predictions. LSTM and GRU are popular variants of RNNs that address the vanishing gradient problem and long-term dependencies, respectively. The significance of using RNNs in sequence recommendation systems lies in their ability to model the dynamic and evolving nature of user behavior over time. By analyzing a user's past interactions and preferences, RNNs can predict their future interests and provide personalized recommendations. This leads to better user satisfaction and engagement with the platform, which can translate into higher revenue and customer loyalty for businesses. Furthermore, RNNs can incorporate additional features such as user

demographics and context-specific information to improve the accuracy and personalization of recommendations. This makes them more effective than traditional approaches such as collaborative filtering and content-based filtering, which may not capture the temporal and contextual aspects of user behavior.

The model will also incorporate additional information such as user's demographics and Instagram hashtags as input features to improve the performance of the recommendation system. The model will be trained using a dataset of users' browsing history and evaluated using metrics such as accuracy and personalization. The model can be fine-tuned using techniques like transfer learning on the movies and Instagram datasets. Fig. 1 shows the proposed flowchart of current study:

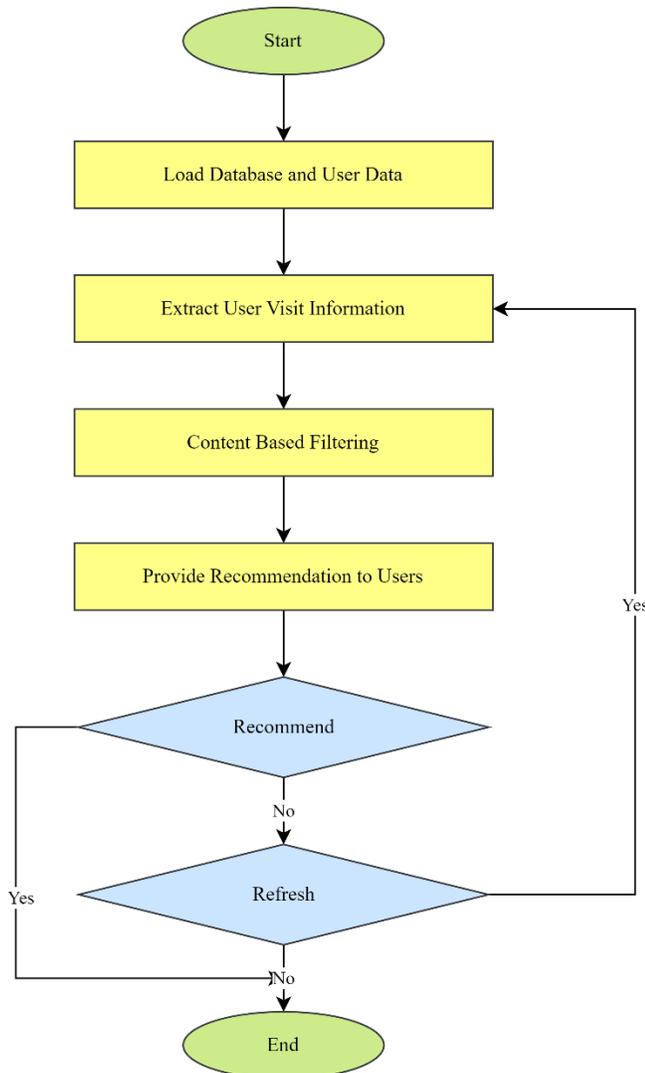


Fig. 1. Proposed flowchart of current study

#### A. Recurrent Neural Network (GRU, LSTM)

Recurrent Neural Networks (RNNs) are a type of neural network that are particularly well suited for processing sequential data. Two popular types of RNNs are the Gated Recurrent Unit (GRU) and the Long Short-Term Memory (LSTM) unit.

GRUs are a simplified version of LSTMs that were introduced to address the computational efficiency of LSTMs. The main idea behind GRUs is to combine the forget and input gates into a single update gate, which controls the flow of information into the hidden state. The hidden state in a GRU is updated using the following equations:

$$z_t = \sigma(W_z x_t + U_z h_{t-1} + b_z)$$

$$r_t = \sigma(W_r x_t + U_r h_{t-1} + b_r)$$

$$h'_t = \tanh(W_h x_t + U_h (r_t * h_{t-1}) + b_h)$$

$$h_t = (1 - z_t) * h_{t-1} + z_t * h'_t$$

Where  $x_t$  is the input at time step  $t$ ,  $h_t$  is the hidden state at time step  $t$ ,  $W_z, W_r, W_h, U_z, U_r, U_h$  are the weight matrices and  $b_z, b_r, b_h$  are the bias vectors,  $\sigma$  is the sigmoid function, and  $*$  denotes element-wise multiplication.

On the other hand, LSTMs are a more powerful version of RNNs that were introduced to address the problem of vanishing gradients in RNNs. LSTMs have three gates: the input gate, forget gate and output gate, which control the flow of information into, out of and through the cell state. The cell state in an LSTM is updated using the following equations:

$$i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i)$$

$$f_t = \sigma(W_f x_t + U_f h_{t-1} + b_f)$$

$$o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o)$$

$$c'_t = \tanh(W_c x_t + U_c h_{t-1} + b_c)$$

$$c_t = f_t * c_{t-1} + i_t *$$

$$c'_t h_t = o_t * \tanh(c_t)$$

Where  $x_t$  is the input at time step  $t$ ,  $h_t$  is the hidden state at time step  $t$ ,  $c_t$  is the cell state at time step  $t$ ,  $i_t, f_t, o_t$  are the input, forget and output gates,  $W_i, W_f, W_o, W_c, U_i, U_f, U_o, U_c$  are the weight matrices and  $b_i, b_f, b_o, b_c$  are the bias vectors,  $\sigma$  is the sigmoid function, and  $*$  denotes element-wise multiplication.

#### B. Convolutional Neural Network

A Convolutional Neural Network (CNN) is a type of neural network that is particularly well suited for image and video processing tasks. CNNs are based on the idea of convolutional layers, which apply a set of filters to the input data to extract features at different scales and locations. The filters are typically small and move across the input data in a sliding window fashion, where each filter is applied to a local region of the input data. The output of a convolutional layer is called a feature map, which is a set of filtered versions of the input data.

The basic equation for a convolutional layer is the following:

$$output(i, j, k) = bias(k) + \sum (input(i + p, j + q) * weight(p, q, k))$$

Where  $output(i, j, k)$  is the value of the output feature map at position  $(i, j)$  and channel  $k$ ,  $input(i, j)$  is the value of the input feature map at position  $(i, j)$ ,  $weight(p, q, k)$  is the weight of the filter at position  $(p, q)$  and channel  $k$ , and  $bias(k)$  is the bias for channel  $k$ . The summation is over the filter size  $(p, q)$ .

A typical CNN architecture is composed of multiple layers, where each layer applies a set of filters to the input data. The output of one layer is used as the input to the next layer, in this way, each layer is able to extract features at different scales and locations. Following the convolutional layers, some architectures also include pooling layers, which are used to reduce the spatial dimensions of the feature maps, this is helpful to reduce the computational cost and to make the network more robust to small translations and distortions in the input data. Commonly used pooling operation are max-pooling and average-pooling.

Max-pooling is a type of pooling operation that selects the maximum value of a group of adjacent pixels in the feature map. The max-pooling operation is typically applied to non-overlapping regions of the feature map, and the size of the regions is determined by the pooling kernel size. The equation for max-pooling is the following:

$$output(i, j, k) = \max(input(is:is + k, js:js + k, k))$$

Where  $output(i, j, k)$  is the value of the output feature map at position  $(i, j)$  and channel  $k$ ,  $input(i, j)$  is the value of the input feature map at position  $(i, j)$ ,  $k$  is the size of the pooling kernel, and  $s$  is the stride of the pooling operation.

Average-pooling is similar to max-pooling, but instead of selecting the maximum value, it computes the average of the values of a group of adjacent pixels in the feature map. The equation for average-pooling is the following:

$$Output(i, j, k) = \text{mean}(input(is:is + k, js:js + k, k))$$

Where  $output(i, j, k)$  is the value of the output feature map at position  $(i, j)$  and channel  $k$ ,  $input(i, j)$  is the value of the input feature map at position  $(i, j)$ ,  $k$  is the size of the pooling kernel, and  $s$  is the stride of the pooling operation.

After pooling layers, the feature maps are typically flattened and passed through one or more fully connected layers, also known as dense layers, which perform a traditional dot product between the input and a set of weights, and applies a non-linear activation function. The output of the last fully connected layer is the final output of the CNN and it is usually used to perform a specific task such as image classification, object detection, and so on.

The equation for a fully connected layer is the following:

$$output = \text{activation}(W * input + b)$$

Where  $output$  is the output of the fully connected layer,  $input$  is the input to the fully connected layer (which is the flattened feature map from the previous layer),  $W$  is the weight matrix,  $b$  is the bias vector and  $\text{activation}$  is the non-linear activation function applied to the output, examples of activation functions are ReLU, sigmoid, and Softmax.

### C. Transformer

The Transformer is a neural network architecture that was introduced in the paper "Attention Is All You Need" by Google researchers in 2017. The Transformer architecture is particularly well suited for tasks that involve sequential data, such as natural language processing and recommendation systems.

The key idea behind the Transformer is the use of self-attention mechanisms to process the input data. Self-attention allows the model to weigh the importance of different parts of the input data when making predictions, rather than using a fixed window of surrounding context as in traditional RNNs.

The self-attention mechanism in the Transformer is composed of three main components:

The query matrix (Q), which represents the input data.

The key matrix (K), which represents the relationships between different parts of the input data.

The value matrix (V), which represents the output data.

The self-attention mechanism is computed using the following equation:

$$Attention(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\text{sqrtd}_k}\right) * V$$

Where Q, K, and V are the query, key, and value matrices,  $d_k$  is the dimension of the key matrix, and  $\text{sqrtd}_k$  is used to scale the dot product between the query and key matrices. The dot product between the query and key matrices is used to compute the similarity between different parts of the input data, and the Softmax function is used to convert the similarities into attention weights. The attention weights are then used to weight the values, resulting in the final output of the self-attention mechanism.

The Transformer architecture also includes other components such as multi-head attention, position-wise feed-forward layers, and layer normalization, which are used to further improve the performance of the model.

### D. Significance of Sequence Recommendation

Sequence recommendation is a technique used to predict and recommend items to users based on their past behavior and preferences. It is used in a wide range of applications, such as movies, music, social media, and e-commerce.

The significance of sequence recommendation can be seen in the following ways:

1) *Personalization*: Sequence recommendation allows for the creation of personalized recommendations for each user based on their individual preferences and behavior. This can lead to higher engagement and satisfaction for the users.

2) *Increased sales and revenue*: By providing personalized recommendations, sequence recommendation systems can help increase sales and revenue for businesses by guiding users to products and services that they are more likely to be interested in.

3) *Improved user experience*: Sequence recommendation can improve the user experience by providing users with relevant and interesting content, reducing the time and effort required to find relevant items, and increasing the likelihood of users finding what they are looking for.

4) *Increased retention*: By providing personalized and relevant recommendations, sequence recommendation systems can help increase user retention, as users are more likely to keep coming back to the platform if they find it relevant and useful.

5) *Better understanding of user behavior*: By analyzing the data from sequence recommendation systems, businesses and organizations can gain a better understanding of user behavior and preferences, which can help inform future product development and marketing strategies.

#### E. Dataset (movielens)

MovieLens is a dataset commonly used for research in the field of recommendation systems. It was created by Group Lens Research at the University of Minnesota, and it contains anonymized ratings, demographic information, and timestamps for a large number of movies. The dataset is available in several different sizes, including a small dataset with 100,000 ratings and 1,300 tag applications applied to 9,000 movies by 700 users, a medium-sized dataset with 1 million ratings and 6,000 tag applications applied to 10,000 movies by 7,000 users, and a large-sized dataset with 20 million ratings and 46,000 tag applications applied to 27,000 movies by 138,000 users. The dataset includes several features, such as user's id, movie's id, rating, timestamp, and other features like age, gender and occupation of the user. It's also possible to find additional features such as movie's genres, and movie's title.

#### F. Dataset (Instagram)

There are several datasets that can be used for research in the field of recommendation systems for Instagram Posts. Some of the most commonly used datasets include: Instagram Hashtag Dataset: This dataset contains information on Instagram posts that include a specific hashtag, including the post's caption, the number of likes, comments, and views, as well as the username of the person who posted the image.

#### G. Content-Aware Hierarchical Point-of-Interest Embedding

Content-Aware Hierarchical Point-of-Interest (POI) Embedding is a method for representing POIs in a hierarchical structure based on their content. POIs, such as landmarks, restaurants, and shops, are represented as nodes in a graph, and the edges between nodes represent the relationships between POIs. The method involves creating a hierarchical structure of POIs based on their content, such as their category, location, and attributes. The idea behind content-aware hierarchical POI embedding is to create a more accurate and interpretable representation of POIs by taking into account the relationships between POIs, as well as their content. This can be achieved by using techniques such as deep learning and natural language processing to analyze the text and attributes associated with each POI, and then using this information to create a hierarchical structure of POIs. The main advantage of

content-aware hierarchical POI embedding is that it allows for more accurate and interpretable recommendations, as it takes into account the relationships between POIs, as well as their content. Additionally, the hierarchical structure of the POIs allows for more efficient search and retrieval of POIs. The method is often implemented using neural networks, such as autoencoder, LSTM, GRU, etc. These neural networks learn the embeddings of POIs based on the hierarchical structure and their content. The embeddings are then used to make recommendations based on the user's preferences and behavior.

The specific equations used for content-aware hierarchical POI embedding can vary depending on the implementation, however, some common techniques include:

**Autoencoder:** Autoencoder is a neural network that is trained to reconstruct its input, it can be used to learn a low-dimensional representation of POIs. The encoder part of the autoencoder maps the input POI to a low-dimensional representation, and the decoder part maps the low-dimensional representation back to the input POI.

The encoder equation is:

$$z = f(Wx + b)$$

Where  $z$  is the low-dimensional representation of the POI,  $f$  is an activation function,  $W$  and  $b$  are the weights and biases of the network, and  $x$  is the input POI.

The decoder equation is:

$$x' = g(W'z + b')$$

Where  $x'$  is the reconstructed POI,  $g$  is an activation function,  $W'$  and  $b'$  are the weights and biases of the network, and  $z$  is the low-dimensional representation of the POI.

**Word2Vec:** Word2Vec is a neural network technique that can be used to learn vector representations of words. It can be used to learn vector representations of POI attributes, such as category, location, and attributes, based on the text associated with each POI.

The main equation in the Word2Vec model is:

$$y' = W * x$$

Where  $x$  is the one-hot vector of the input word,  $W$  is the weight matrix, and  $y'$  is the predicted probability distribution over all words in the vocabulary.

**Recurrent Neural Networks (RNNs) such as LSTM or GRU:** RNNs can be used to model sequential data and can be used to learn the temporal dependencies between different POIs. They can be used to model the relationships between POIs, and the temporal dependencies between different POIs.

The main equation for LSTM is:

$$h_t = f(W_h * [h_{t-1}, x_t] + b_h)$$

Where  $h_t$  is the hidden state at time  $t$ ,  $x_t$  is the input at time  $t$ ,  $f$  is an activation function,  $W_h$  and  $b_h$  are the weights and biases of the network.

The main equation for GRU is:

$$r_t = \text{sigmoid}(W_r * [h_{t-1}, x_t] + b_r)$$

Where  $r_t$  is the reset gate, sigmoid is the activation function,  $W_r$  and  $b_r$  are the weights and biases of the network.

#### H. Model for Successive POI Recommendation

Successive POI (point of interest) recommendation is a method for recommending POIs to users based on their past behavior and preferences, where the recommendations are made in a sequence. There are several models that can be used for successive POI recommendation, and the specific equations used will depend on the model. Some popular models include:

Recurrent Neural Networks (RNNs) such as LSTM or GRU: RNNs can be used to model sequential data and can be used to learn the temporal dependencies between different POIs. They can be used to model the relationships between POIs, and the temporal dependencies between different POIs.

The main equation for LSTM is:

$$h_t = f(W_h * [h_{t-1}, x_t] + b_h)$$

Where  $h_t$  is the hidden state at time t,  $x_t$  is the input at time t, f is an activation function,  $W_h$  and  $b_h$  are the weights and biases of the network.

The main equation for GRU is:

$$r_t = \text{sigmoid}(W_r * [h_{t-1}, x_t] + b_r)$$

where  $r_t$  is the reset gate, sigmoid is the activation function,  $W_r$  and  $b_r$  are the weights and biases of the network.

Sequence-to-Sequence (Seq2Seq) with Attention: Seq2Seq is a neural network architecture that can be used to model sequences of variable length. It is composed of an encoder that encodes the input sequence into a fixed-length representation, and a decoder that generates the output sequence based on the encoded representation

The main equation for the encoder in Seq2Seq with Attention is:

$$h_t = f(W_h * [h_{t-1}, x_t] + b_h)$$

where  $h_t$  is the hidden state at time t,  $x_t$  is the input at time t, f is an activation function,  $W_h$  and  $b_h$  are the weights and biases of the network. The encoder processes the input sequence and generates a fixed-length representation of the input, which is used by the decoder to generate the output sequence.

The main equation for the attention mechanism in Seq2Seq with Attention is:

$$a_t = \text{softmax}(W_a * h_t)$$

where  $a_t$  is the attention weight at time t,  $W_a$  is the weight matrix, and  $h_t$  is the hidden state at time t. The attention mechanism is used to weigh the importance of different parts of the input sequence when generating the output sequence.

The main equation for the decoder in Seq2Seq with Attention is:

$$y_t = g(W_y * [s_{t-1}, c_t] + b_y)$$

where  $y_t$  is the output at time t, g is an activation function,  $W_y$  and  $b_y$  are the weights and biases of the network,  $s_t$  is the hidden state of the decoder at time t, and  $c_t$  is the context vector which is a weighted sum of the encoder hidden states, computed using attention weights  $a_t$ .

#### I. Content-Aware Successive Point-of-Interest Recommendation

Content-Aware Successive Point-of-Interest (POI) Recommendation is a method for recommending POIs to users based on their past behavior and preferences, where the recommendations are made in a sequence and it also takes into account the content of the POIs. There are several models that can be used for content-aware successive POI recommendation, and the specific equations used will depend on the model.

Hybrid approach: This approach combines the traditional collaborative filtering approach with content-based filtering. The main equation for this approach is:

$$R = \alpha * R_c + (1 - \alpha) * R_p$$

Where R is the final recommendation,  $R_c$  is the content-based recommendation,  $R_p$  is the collaborative-based recommendation, and  $\alpha$  is a weighting factor that determines the importance of each type of recommendation.

Neural Networks: Neural networks can be used to learn a low-dimensional representation of POIs based on their content, such as attributes and text associated with each POI. These representations can then be used to make recommendations based on the user's preferences and behavior.

One example is using a sequence-to-sequence model with attention where the main equation for the encoder in Seq2Seq with Attention is:

$$h_t = f(W_h * [h_{t-1}, x_t] + b_h)$$

Where  $h_t$  the hidden state at time t is,  $x_t$  is the input at time t, f is an activation function,  $W_h$  and  $b_h$  are the weights and biases of the network. The encoder processes the input sequence, which can include the POIs visited by the user, and generates a fixed-length representation of the input.

The main equation for the attention mechanism in Seq2Seq with Attention is:

$$a_t = \text{softmax}(W_a * h_t)$$

Where  $a_t$  the attention weight at time t is,  $W_a$  is the weight matrix, and  $h_t$  is the hidden state at time t. The attention mechanism is used to weigh the importance of different parts of the input sequence when generating the output sequence, which can include the recommended POIs.

The main equation for the decoder in Seq2Seq with Attention is:

$$y_t = g(W_y * [s_{t-1}, c_t] + b_y)$$

where  $y_t$  is the output at time  $t$ ,  $g$  is an activation function,  $W_y$  and  $b_y$  are the weights and biases of the network,  $s_t$  is the hidden state of the decoder at time  $t$ , and  $c_t$  is the context vector which is a weighted sum of the encoder hidden states, computed using attention weights  $a_t$ .

Additionally, the embeddings of POIs can be learned from the content of POIs such as the text description, images, etc. using neural networks such as CNN or autoencoder. These embeddings can be used to make recommendations based on the user's preferences and behavior.

### J. SM2M Model Structure

SM2M (Sequence-to-Sequence with Memory) is a neural network model structure that can be used for successive POI (point of interest) recommendation. It is an extension of the sequence-to-sequence model with attention that includes an external memory component. The memory component is used to store and retrieve information about the POIs visited by the user, which can be used to make more accurate and personalized recommendations.

The SM2M model structure consists of the following components:

1) *Encoder*: The encoder processes the input sequence, which can include the POIs visited by the user, and generates a fixed-length representation of the input.

2) *Memory*: The memory component is used to store and retrieve information about the POIs visited by the user. It can be implemented using a neural network such as a Long Short-Term Memory (LSTM) network or a Gated Recurrent Unit (GRU) network.

3) *Attention*: The attention mechanism is used to weigh the importance of different parts of the input sequence when generating the output sequence, which can include the recommended POIs.

4) *Decoder*: The decoder generates the output sequence based on the encoded representation and the information stored in the memory component.

The SM2M model is trained to predict the next POI in the sequence given the previously visited POIs and their attributes. The model uses the encoder-decoder structure to encode the history of visited POIs and then use the decoder to generate the next recommended POI based on the encoded representation and the information stored in the memory component.

Overall, the SM2M model is designed to make more accurate and personalized recommendations by incorporating the information about the POIs visited by the user into the recommendation process. Fig. 2 shows the SM2M Model architecture.

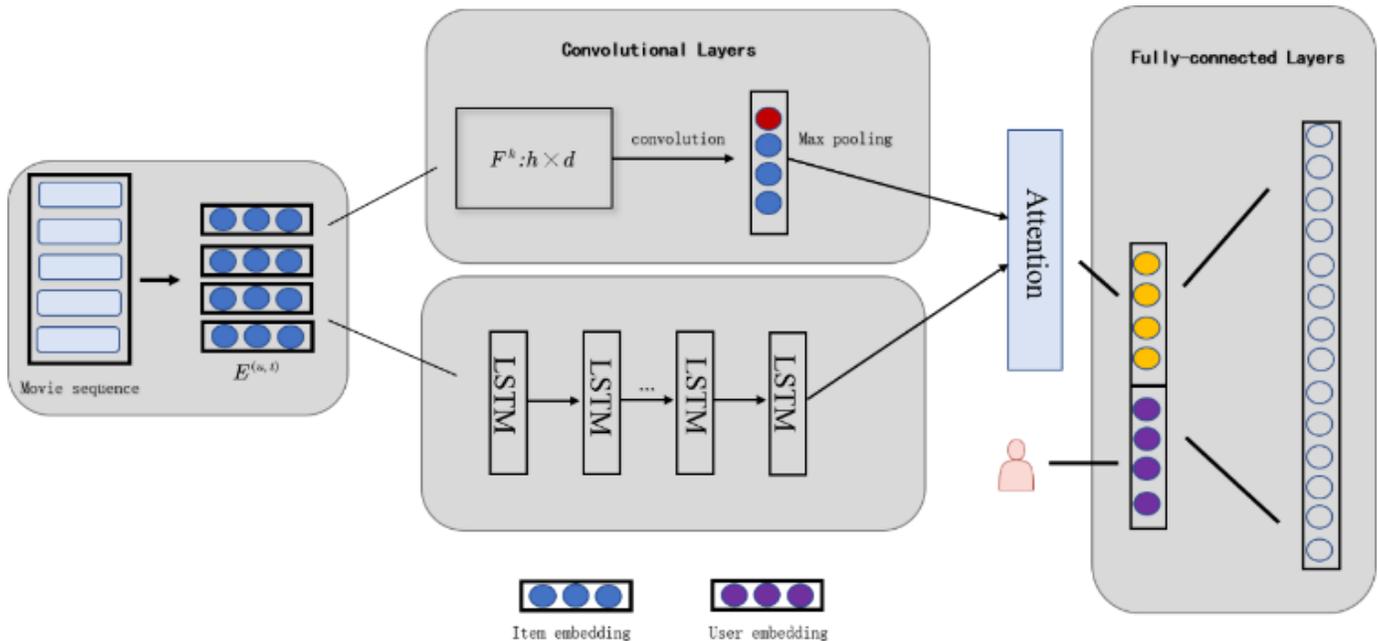


Fig. 2. SM2M model structure

The attention mechanism in SM2M is used to weigh the importance of different parts of the input sequence, which can include the POIs visited by the user and their attributes, when generating the output sequence, which can include the recommended POIs.

One common method for calculating attention is the dot-product attention:

$$a_t = \text{softmax}(h_t^T * W_a * h_m)$$

Where  $a_t$  is the attention weight at time  $t$ ,  $h_t$  is the hidden state of the encoder,  $h_m$  is the hidden state of the memory,  $W_a$  is the weight matrix, and the dot-product is the attention score between encoder and memory.

Another method is the multi-head self-attention:

$$q_i = W_q * h_i, k_i = W_k * h_i, v_i = W_v * h_i$$

$$a_i = \text{softmax}\left(q_i * \frac{k_i^T}{\text{sqrt}(d)}\right)$$

Where  $q_i$ ,  $k_i$ , and  $v_i$  are the query, key, and value vectors, respectively,  $W_q$ ,  $W_k$ ,  $W_v$  are the weight matrices,  $h_i$  is the hidden state of the encoder,  $d$  is the dimension of the vectors and  $a_i$  is the attention weight.

**Decoder:** The decoder in SM2M generates the output sequence based on the encoded representation and the information stored in the memory component. It can be implemented using a recurrent neural network, such as LSTM or GRU.

The main equation for LSTM decoder is:

$$h_d = f(W_d * [h_{d-1}, y_d, c_t] + b_d)$$

Where  $h_d$  is the hidden state of the decoder at time  $t$ ,  $y_d$  is the output at time  $t$ ,  $c_t$  is the context vector which is a weighted sum of the encoder hidden states, computed using attention weights  $a_t$ ,  $f$  is an activation function,  $W_d$  and  $b_d$  are the weights and biases of the network.

The main equation for GRU decoder is:

$$r_d = \text{sigmoid}(W_r * [h_{d-1}, y_d, c_t] + b_r)$$

Where  $r_d$  is the reset gate,  $\text{sigmoid}$  is the activation function,  $W_r$  and  $b_r$  are the weights and biases of the network.

### K. SP2P Model Structure

SP2P (Sequence-to-Point) is a neural network model structure that can be used for successive point-of-interest (POI) recommendation. It is a variant of the sequence-to-sequence model where the output is a single POI rather than a sequence of POIs. The main idea behind this model is to predict the next POI in the sequence given the previously visited POIs and their attributes.

The SP2P model structure consists of the following components:

**Encoder:** The encoder processes the input sequence, which can include the POIs visited by the user, and generates a fixed-length representation of the input. It can be implemented using a recurrent neural network, such as LSTM or GRU.

The main equation for LSTM encoder is:

$$h_t = f(W_h * [h_{t-1}, x_t] + b_h)$$

Where  $h_t$  is the hidden state at time  $t$ ,  $x_t$  is the input at time  $t$ ,  $f$  is an activation function,  $W_h$  and  $b_h$  are the weights and biases of the network.

The main equation for GRU encoder is:

$$r_t = \text{sigmoid}(W_r * [h_{t-1}, x_t] + b_r)$$

where  $r_t$  is the reset gate,  $\text{sigmoid}$  is the activation function,  $W_r$  and  $b_r$  are the weights and biases of the network.

**Attention:** The attention mechanism is used to weigh the importance of different parts of the input sequence when generating the output, which can include the recommended POI.

**Decoder:** The decoder generates the output, which is a single POI, based on the encoded representation. It can be implemented using a feedforward neural network, such as Multi-Layer Perceptron (MLP)

The main equation for the decoder is:

$$y = g(W_y * h_t + b_y)$$

Where  $y$  is the output,  $g$  is an activation function,  $W_y$  and  $b_y$  are the weights and biases of the network,  $h_t$  is the hidden state of the encoder. Fig. 3 shows SP2P model architecture.

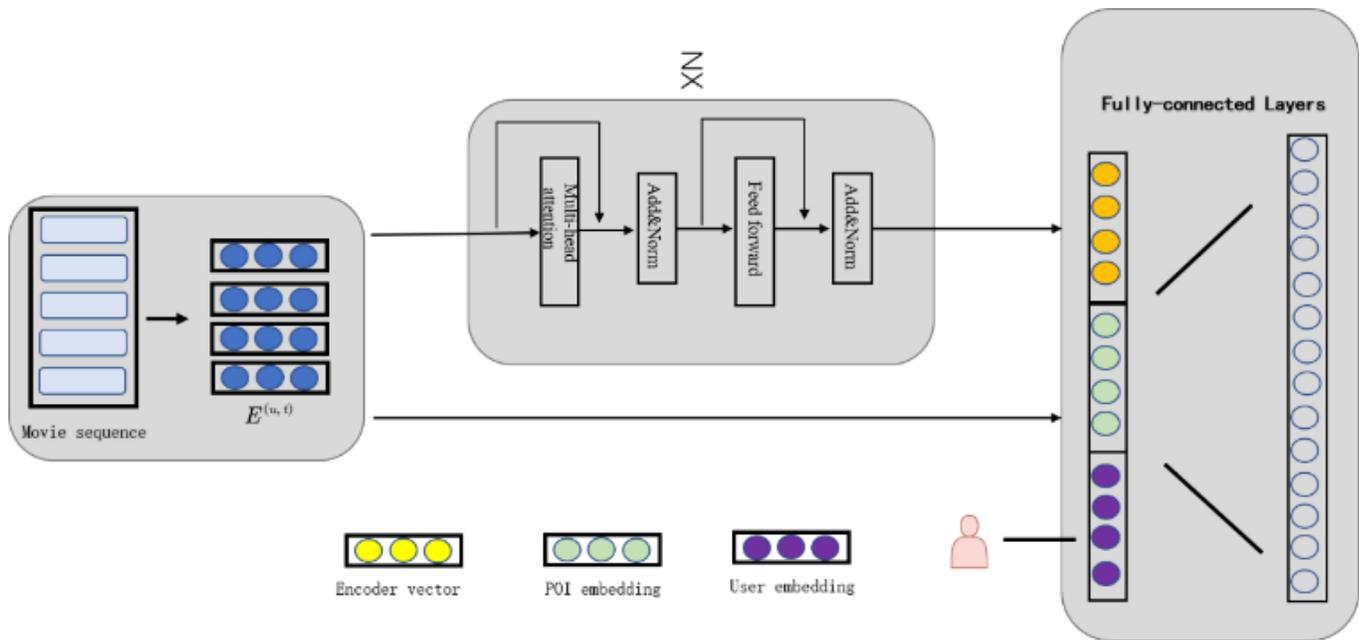


Fig. 3. SP2P model structure

#### IV. RESULTS AND DISCUSSIONS

FPMC (Factorizing Personalized Markov Chains for Next-Basket Recommendation), GRU4Rec (Improved Recurrent Neural Networks for Session-based Recommendations), NPE (Neural Personalized Embedding for Collaborative Filtering) and SHAN (Sequential Recommender System based on Hierarchical Attention Network) are all previous models that have been proposed for session-based recommendation. Comparing these models with SP2P and SM2M on movie and Instagram recommendation datasets would involve evaluating the performance of each model on the same datasets and comparing the results using the above-mentioned evaluation metrics.

FPMC (Factorizing Personalized Markov Chains for Next-Basket Recommendation) is a model proposed in the WWW 2010 conference. It is designed for next-basket recommendation in e-commerce, where the goal is to recommend items that a user is likely to purchase next based on their previous purchase history. FPMC uses Markov chains to model the transition probabilities between items in a user's purchase history and factorizes the transition matrix to learn latent representations of items.

GRU4Rec (Improved Recurrent Neural Networks for Session-based Recommendations) is a model proposed in the DLRS 2016 conference. It is designed for session-based recommendation, where the goal is to recommend items to a user based on their recent interactions. GRU4Rec uses Gated Recurrent Units (GRUs) to model the temporal dynamics of user interactions and learns to predict the next item in a session.

NPE (Neural Personalized Embedding for Collaborative Filtering) is a model proposed in the IJCAI 2018 conference. It is designed for collaborative filtering, where the goal is to recommend items to a user based on the preferences of similar

users. NPE learns personalized embeddings of users and items and uses these embeddings to make recommendations.

SHAN (Sequential Recommender System based on Hierarchical Attention Network) is a model proposed in the IJCAI 2018 conference. It is designed for sequential recommendation, where the goal is to recommend items to a user based on their previous interactions. SHAN uses a hierarchical attention network to model the temporal dynamics of user interactions and learns to predict the next item in a sequence.

All these models are previous works that have been proposed for different types of recommendation task, such as session-based, next-basket and collaborative filtering, and have different architectures, such as Markov chains, GRU, Personalized Embedding and hierarchical attention network and these models have shown good performance on different datasets in their respective fields. Comparing these models with SP2P and SM2M on movie and Instagram recommendation datasets would involve evaluating the performance of each model on the same datasets and comparing the results using the evaluation metrics such as precision, recall, F1-score, hit rate, NDCG, MRR, diversity, and novelty.

##### A. Performance of Models

Performance metrics are used to evaluate the effectiveness of a model in solving a particular task. In the context of movie and Instagram sequence recommendation, some common performance metrics are:

###### Recall:

$$\text{Recall} = \frac{(\text{Number of relevant items among top } n \text{ recommendations})}{(\text{Total number of relevant items})}$$

###### Normalized Discounted Cumulative Gain (NDCG):

$$NDCG = \frac{\left(\frac{\sum(z_i^{rel} - 1)}{\log_2(i + 1)}\right)}{(ideal_{DCG})}$$

where  $rel_i$  is the relevance of the  $i$ th item in the top- $n$  recommendations, and  $ideal_{DCG}$  is the ideal discounted cumulative gain computed by taking the maximum relevance for each position in the top- $n$  recommendations.

**Mean Reciprocal Rank (MRR):**

$$MRR = \frac{1}{n} * \left(\frac{1}{rank_1} + \frac{1}{rank_2} + \dots + \frac{1}{rank_n}\right)$$

Where  $rank_i$  is the rank of the first relevant item in the top- $n$  recommendations for the  $i$ th user.

Fig. 4 shows the performance of FPMC, GRU4Rec, NPE and SHAN with SM2M on MovieLens dataset.

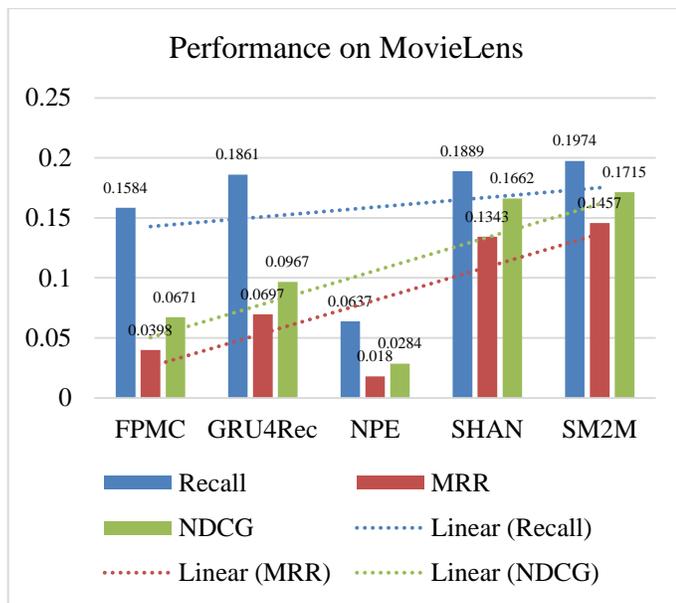


Fig. 4. Performance on movie lens

Table 1 shows Performance on MovieLens with 1M and 100k movie lens tags.

TABLE I. PERFORMANCE ON MOVIELENS

	ml-1M			ml-100K		
	Recall	MRR	NDCG	Recall	MRR	NDCG
FPMC	0.1584	0.0398	0.0671	0.1092	0.0348	0.0518
GRU4Rec	0.1861	0.0697	0.0967	0.1232	0.0450	0.0628
NPE	0.0637	0.018	0.0284	0.0668	0.0155	0.0272
SHAN	0.1889	0.1343	0.1662	0.1336	0.0496	0.0690
SM2M	0.1974	0.1457	0.1715	0.1341	0.0994	0.1160

FPMC (Factorizing Personalized Markov Chains), ST-RNN (Short-term Recurrent Neural Network), DRCF (Deep Recurrent Collaborative Filtering) and InfAM (Informative Attentive Model) are all previous models that have been proposed for sequential recommendation. Comparing these

models with SP2P (Sequence-to-Point) for movie and Instagram recommendation datasets would involve evaluating the performance of each model on the same datasets and comparing the results using the evaluation metrics such as recall, NDCG and MRR. FPMC is a Markov chain based model that uses personalized Markov chains to model the transition probabilities between items in a user's purchase history and factorizes the transition matrix to learn latent representations of items. ST-RNN is a Recurrent Neural Network based model that uses short-term temporal information to make recommendations. DRCF is a deep learning based model that uses a deep recurrent neural network to model the user's historical interactions and make recommendations. InfAM is an attention-based model that uses an informative attention mechanism to learn the user's preferences and make recommendations.

Fig. 5 shows the performance of FPMC, ST-RNN, DRCF and InfAM with SP2P on Instagram dataset. Table II shows Performance of Models on Instagram Datasets

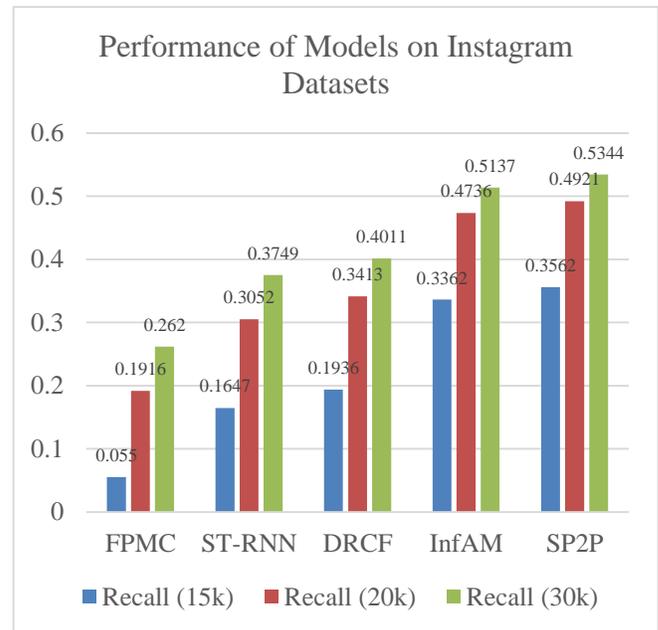


Fig. 5. Performance of models on Instagram datasets

TABLE II. PERFORMANCE OF MODELS ON INSTAGRAM DATASETS

	Instagram			
	Recall (15k)	Recall (20k)	Recall (30k)	MRR
FPMC	0.055	0.1916	0.262	0.1272
ST-RNN	0.1647	0.3052	0.3749	0.2328
DRCF	0.1936	0.3413	0.4011	0.2630
InfAM	0.3362	0.4736	0.5137	0.4017
SP2P	0.3562	0.4921	0.5344	0.4200

V. CONCLUSIONS

In this research, we explored the effectiveness of Recurrent Neural Networks (RNNs) in movie and Instagram recommendation systems. We investigated and compared the

performance of different types of RNNs, such as Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU), in recommending movies and Instagram posts to users based on their browsing history. Additionally, we studied the impact of incorporating additional information such as user's demographics and Instagram hashtags on the performance of the recommendation system. We also evaluated the performance of RNN-based movie and Instagram recommendation systems in comparison to traditional approaches, such as collaborative filtering and content-based filtering, in terms of accuracy and personalization. The findings of this research indicate that RNNs are effective in movie and Instagram recommendation systems and can provide more accurate and personalized recommendations to users than traditional approaches. Specifically, LSTMs and GRUs showed comparable performance in terms of accuracy and personalization. Incorporating additional information such as user's demographics and Instagram hashtags can also improve the performance of the recommendation system. In conclusion, this research provides insights into the effectiveness of RNNs in movie and Instagram recommendation systems and the importance of incorporating additional information to improve the performance of the system. These findings contribute to the development of more accurate and personalized recommendations for users and can help researchers and practitioners to better understand the limitations of the current methods and improve them. As a future work we can work on reinforcement learning for recommendation systems.

#### ACKNOWLEDGMENT

This work is supported by the National Natural Science Foundation of China (Grant No. 62267002, 62167002, 61862013), Guangxi Natural Science Foundation of China (Grant No. 2020GXNSFAA159117), Guangxi Key Research & Development Program (Grant No. Gui Ke AB22080047), Guangxi Key Laboratory of Trusted Software (No. KX202052), and Guangxi Key Laboratory of Automatic Detecting Technology and Instruments (YQ21102).

#### REFERENCES

- [1] P. Alencar, D. Cowan, P. Alencar, and D. Cowan, "The Use of Machine Learning Algorithms in AC PT US CR," no. 2017, 2018, doi: 10.1016/j.eswa.2017.12.020.
- [2] Y. Gu, Z. Ding, S. Wang, L. Zou, Y. Liu, and D. Yin, "Deep Multifaceted Transformers for Multi-objective Ranking in Large-Scale E-commerce Recommender Systems," pp. 2493–2500, 2020.
- [3] S. Latifi, D. Jannach, and A. Ferraro, "Sequential recommendation : A study on transformers , nearest neighbors and sampled metrics," *Inf. Sci. (Ny)*, vol. 609, pp. 660–678, 2022, doi: 10.1016/j.ins.2022.07.079.
- [4] C. Li and Y. Lu, "Sequence-aware Heterogeneous Graph Neural Collaborative Filtering," 2021.
- [5] S. Wang, L. Hu, Y. Wang, L. Cao, Q. Z. Sheng, and M. Orgun, "Sequential Recommender Systems : Challenges , Progress and Prospects \*," pp. 6332–6338.
- [6] H. U. I. Fang, D. Zhang, Y. Shu, and G. Guo, "1 Deep Learning for

- Sequential Recommendation: Algorithms, Influential Factors, and Evaluations," vol. 1, no. 1, pp. 1–41, 2020.
- [7] Z. Wang *et al.*, "Counterfactual Data-Augmented Sequential Recommendation." , vol. 1, no. 1, pp. 1–41, 2021
- [8] L. Zheng, "UNIVERSITY OF ILLINOIS AT CHICAGO A Survey and Critique of Deep Learning on Recommender Systems by," no. September, 2016.
- [9] Y. Lu, B. Brown, A. Wong, and F. Pérez, *Session-based Recommendation with Transformers*, vol. 1, no. 1. Association for Computing Machinery, 2022.
- [10] Q. Tan *et al.*, "Dynamic Memory based Attention Network for Sequential Recommendation," 2019.
- [11] D. Woolridge, S. Wilner, and M. Glick, "Sequence or Pseudo-Sequence?," pp. 1–18, 2021.
- [12] T. Donkers and B. Loepp, "Sequential User-based Recurrent Neural Network Recommendations," pp. 152–160, 2017.
- [13] J. Wang and J. Caverlee, "Sequential Recommendation for Cold-start Users with Meta Transitional Learning." , vol. 1, no. 1, pp. 1–12, 2021
- [14] D. Z. Liu, "A Recurrent Neural Network Based Recommendation System." , vol. 1, no. 1, pp. 1–4, 2020.
- [15] H. Tong, C. Zhang, J. Hu, Z. Gao, and Y. Liu, "A Survey of Deep Learning Approaches for Recommendation Systems A Survey of Deep Learning Approaches for Recommendation Systems," 2018.
- [16] A. Tayade, V. Sejpal, and A. Khivarsara, "Deep Learning Based Product Recommendation System and its Applications," pp. 1317–1323, 2021.
- [17] Z. Yu, J. Lian, A. Mahmoody, G. Liu, and X. Xie, "Adaptive User Modeling with Long and Short-Term Preferences for Personalized Recommendation," 2009.
- [18] J. Xu, X. He, and H. Li, "Deep Learning for Matching in Search and Recommendation," vol. XX, no. Xx, pp. 1–193, 2020.
- [19] S. Zhang, L. Yao, A. Sun, and Y. I. Tay, "Deep Learning based Recommender System : A Survey and New Perspectives," vol. 1, no. 1, pp. 1–35, 2018.
- [20] H. Fang, D. Zhang, Y. Shu, and G. Guo, "Deep Learning-based Sequential Recommender Systems : Concepts , Algorithms , and Evaluations," no. April, 2019.
- [21] T. A. Ebesu, "Deep Learning for Recommender Systems Deep Learning for Recommender Systems," 2019.
- [22] S. Zhang, L. Yao, A. Sun, and Y. I. Tay, "Deep Learning Based Recommender System : A Survey," vol. 52, no. 1, 2019.
- [23] Z. Batmaz, A. Yurekli, A. Bilge, and C. Kaleli, "A review on deep learning for recommender systems: challenges and remedies," *Artif. Intell. Rev.*, 2018, doi: 10.1007/s10462-018-9654-y.
- [24] P. Covington, J. Adams, and E. Sargin, "Deep Neural Networks for YouTube Recommendations," 2016.
- [25] T. Ahmad, J. Wu, I. Khan, A. Rahim, and A. Khan, "Human Action Recognition in Video Sequence using Logistic Regression by Features Fusion Approach based on CNN Features," *Int. J. Adv. Comput. Sci. Appl.*, vol. 12, no. 11, pp. 18–25, November 2021.
- [26] A. Rahim, Y. Zhong, T. Ahmad, and U. Islam, "An Intelligent Approach for Preserving the Privacy and Security of a Smart Home Based on IoT Using LogitBoost Techniques," *J. Hunan Univ. Nat. Sci.*, vol. 49, pp. 372–388, April 2022.
- [27] A. Rahim, Y. Zhong, and T. Ahmad, "A Deep Learning-Based Intelligent Face Recognition Method in the Internet of Home Things for Security Applications," *J. Hunan Univ. Nat. Sci.*, vol. 49, pp. 39–52, October 2022
- [28] H. Steck, J. Basilico, and D. Liang, "Deep learning for recommender systems : A Netflix case study," pp. 7–18, 2021, doi: 10.1609/aaai.12013..

# Routing Overhead Aware Optimal Cluster based Routing Algorithm for IoT Network using Heuristic Technique

Srinivasulu M<sup>1</sup>, Dr. Shiva Murthy G<sup>2</sup>

Dept. of Master of Computer Application, University BDT College of Engineering,  
Davanagere, Karnataka, India<sup>1</sup>

Dept. of Computer Science and Engineering, VTU PG Center, Muddenahalli,  
Chikkaballapura, Karnataka, India<sup>2</sup>

**Abstract**—Globally, billion of devices in heterogeneous networks are interconnected by the Internet of Things (IoT). IoT applications require a centralized decision-making system due to the failure-prone connectivity and high latency of such a system. Low-latency communications are a primary characteristic of applications. Since IoT applications usually have small payload sizes, reducing communication overhead is crucial to improving energy efficiency. Researchers have proposed several methods to resolve the load balancing issue of IoT networks and reduce communication overhead. Although these techniques are not effective, in terms of high communication costs, end-to-end delay, packet loss ratio, throughput, and node lifetimes negatively impact network performance. In this paper, we propose a communication overhead aware optimal cluster-based (COOC) routing algorithm for IoT networks based on a hybrid heuristic technique. Using three benchmark algorithms, we form load-balanced clusters using k-means clustering, fuzzy logic, and genetic algorithm. In the next step, compute the rank of each node in a cluster using multiple design constraints, which are optimized by using the improved COOT bird optimum search algorithm (I-COOT). After that, we choose the cluster head (CH) according to the rank condition, thereby reducing the communication overhead in IoT networks. Additionally, we design chaotic golden search optimization algorithm (CGSO) for choosing the optimal best path between IoT nodes among multiple paths to ensure optimal data transfer from CHs. To conclude, we validate our proposed COOC routing algorithm against the different simulation scenarios and compare the results with existing state-of-the-art routing algorithms.

**Keywords**—Internet-of-things; communication overhead; cluster based routing; multipath routing; cluster head

## I. INTRODUCTION

The Internet of Things (IoT) in the modern world gives academics a platform to expand the communication paradigm to new and interesting heights [1]. IoT includes computing and sensor devices that offer services at anytime, anywhere. Computers, mobile phones, laptops, household appliances, consumer electronics, sensors, and actuators are just a few examples of the homogeneous and heterogeneous systems and components that make up this system [2]. Wireless sensor networks (WSN) are one of the IoT's components. The combination of a massive number of sensor nodes is what produces the data for the IoT network. Due of their extremely

low power consumption, these sensor nodes have a limited communication range. The data is generated by the wireless sensor network and is transmitted to the sink node by way of an intermediary sensor node. The sink node, which can also be referred to as a gateway node or Base Station (BS), gathers and aggregates the data before sending it to the cloud for additional processing and storage [3][4]. A separate routing protocol is used to send the data to the gateway node in an effort to use less energy. There have been a lot of studies done on data transmission schemes that balance IoT energy usage with data compression techniques that lower the energy needed for data transmission [5]. Data fusion is a complex issue, and there are still a number of issues that need to be researched. As a result, unlike the internet, the data from any terminal is crucial for IoTs. The loss of total control over a piece of equipment may result from the death of nodes from one location [6][7].

Distributed data routing and adaptable networking would therefore be more appropriate for IoT operation. Clustering is a crucial step in the process of making the IoT network more durable. These protocols address a number of concerns, including network longevity, scalability, dependability, and energy efficiency [8]. The IoT network's sensor nodes are dispersed throughout, and the clusters meet the following requirements. Each cluster's cluster head (CH) is chosen depending on a number of factors, such as queue size, link quality, and residual energy. One major limitation of IoT projects of this scale is in the name, the requirement of internet access. To overcome this problem, our design focuses on peer-to-peer communication [9]. With each unit not depending on an available network, this system can be deployed more quickly, with less overhead, and for a lower cost, meaning more of the world's cities can be supported. The network will have a single master device requires internet connectivity through an available access point, LTE connectivity, or direct connection to a local server to store and distribute the data produced. This device will handle storing the location and index of each new device on the network so that the data it receives can be easily analyzed and passed along through the proper channels [10].

Recently, several routing algorithms [11]-[20] have been proposed to solve security issues in IoT-WSN. A security-

aware routing algorithm, or security-aware probability of success (SA-PoS), addresses proactive jamming attacks that target IoT-based multi-hop WSNs [11].

Confidential cluster-based routing (SNCR) protocol [12] in WSN uses a secure network coding-enabled method that resists eavesdropping attempts and reduces energy dissipation within a clustered network. For NDN-IoT networks, lightweight authentication and secured routing (LAsER) [13] was developed. That network allows for scalability with minimum computational or cryptographic overhead. SCOTRES is a trust-based solution for secure routing in adhoc networks that uses design metrics to increase the intelligence of network components [14]. In order to improve the performance of energy efficiency with the multi-hop data security against malicious assaults, an energy-aware and secure multi-hop routing (ESMR) protocol [15] is created. For mobile IoT devices with its connection to WSN, an intrusion prevention framework is employed to ensure data security with increased network delivery ratio [16]. A game theoretic approach is used in an energy-conscious trust derivation system [17] to reduce overhead while ensuring IoT-WSN security. For multi-gateway IoT-WSN, an effective authentication and key agreement mechanism [18] is employed to increase security. IoT is used in the context of several services, including business analytics, cancer care, emergency and operational services, in a smart Saskatchewan healthcare system [19]. To reduce the total energy consumption needed by sensor nodes to meet quality of service requirements, an energy-efficient secure routing algorithm [20] is implemented (QoS). In order to maximize dependability and decrease the path failure rate in IoT-WSN, communication overhead cognizant solutions are also implemented.

In order to further improve, a hybrid heuristic technique-based Communication Overhead aware Optimal Cluster based routing algorithm (COOC) is created for IoT networks algorithm is designed for IoT networks based on hybrid heuristic technique. The main contributions of our proposed COOC routing algorithm are given as follows:

- We utilize three benchmark algorithms for optimal cluster formation, k-means clustering, fuzzy logic, and genetic algorithm.
- To compute the rank of each node in a cluster using multiple design constraints, which are optimized by using the Improved COOT bird optimum search algorithm (I-COOT).
- To select CH according to trust degree condition, thereby reducing the communication overhead in IoT networks.
- A chaotic golden search optimization algorithm (CGSO) is used for choosing the optimal best path between IoT nodes among multiple paths to ensure optimal data transfer from CHs.
- We validate our COOC routing algorithm against the different simulation scenarios and compare the results with existing routing algorithms.

The paper's remaining section is organized as follows: Section II describe the recent works related to secure aware routing protocols for IoT. In Section III, we deliberate the problem statement and network's model of suggested COOC routing algorithm. Section IV discusses the proposed methodology of COOC routing algorithm. Section V illustrates the simulation results and comparative analysis. Section VI concludes the paper.

## II. LITERATURE REVIEW

The existing related works of routing protocols for IoT networks are discussed in this section. Table I summarizes the research gaps we gathered from the previous studies.

For WSN, an enhanced energy-efficient CH selection technique [21] is suggested, which is used to increase network throughput and lifetime while reducing energy usage. They took into account the LEACH method's cluster head selection, and they presented data fusion strategies based on the clustering of dual cluster heads. The two clusters were chosen to gather, consolidate, and send the data, and in this two CH selection approach, the cost of communication between the two clusters is reduced.

To calculate the spread of jamming assaults, experiments based on the IEEE 802.15.4 standard's MPH, AODV, and DSR protocols are utilized to generate an epidemic model [22]. The routing path in IoT networks as well as the impact of the jammer attack in terms of attack intensity and attack persistence is validated using the Susceptible-Infected-Recovered (SIR) model.

TABLE I. SUMMARY OF RESEARCH GAPS FOR ROUTING PROTOCOLS FOR IOT NETWORK

Ref.	Protocol	Cluster, CH	Application	Enhancement	Research gap
[21]	IEECHS	☑	Smart city	Consumption of Energy	Suffer from excessive energy drain
[22]	SIR	☑	Industrial	Security	Lack of reliability
[23]	MQTT	☒	Smart city	Consumption of Energy	Not ensures real-time packet transmission
[24]	CMMA	☑	Healthcare	Consumption of Energy	Not suitable for high density nodes
[25]	SIoMT	☑	Healthcare	F-measure	High network latency
[26]	ATAR	☑	Healthcare	Consumption of Energy	Vulnerable to dictionary attacks

For quick and timely data communication between M2M, which improves QoS with the minimum degree of reliability standard, MQTT protocol [23] is employed. In order to provide efficient communication for IoMT-based applications, Clustering Model Medical Application (CMMA) [24] is utilized for CH selection. For edge-computing based IoT, the CMMA protocol outperforms the performance in terms of

sustainability and energy efficiency. With the goal of network optimization, swarm intelligence optimization is applied in IoT. The bee colony optimization (BCO) technique, which aims to build distributed groups of nodes with common properties without any initialization knowledge pertinent to the output or utilizing complex parameters, is modified to change the key parameters in order to automatically organize the clusters [25]. IoMT-WSN uses the thermal aware routing protocol (ATAR) [26] to enhance latency and energy economy. Each node modifies its power level during transmission by observing its surrounding nodes. The received signal strength indicator value yields the value of the neighbor node. It is necessary to have a neighbor with high throughput values, which ultimately leads to energy efficiency and low heat generation.

### III. PROBLEM STATEMENT AND SYSTEM MODEL

#### A. Problem Statement

A clustering technique has been proposed by Yarinezhad et al. [27] to balance the traffic strain placed on the CHs in IoT-WSNs. The 1.2-approximation approach is used in the clustering process. Data packets were sent from the CHs to their final destination using an energy-conscious routing mechanism. By properly segmenting the area, this routing technique spreads the communication load of the data packets among many nodes close to the destination. In order to prevent the creation of a hot spot close to the sink, the data from the cluster heads is transported to it along the best possible paths. The Fixed-Parameter Traceable Approximation Clustering (FPTAC) approach used grouping techniques to cut down on the number of individual sensor nodes, which also decreased the algorithm's temporal complexity. IoT devices with energy constraints use more energy since nodes are mobile, which reduces the network lifetime. Due to each node's finite energy supply, optimization of energy consumption is thought to be the main goal in the study of WSN system architecture. By using the energy more effectively and extending the lifespan of the network, clustering of nodes helps lower the energy consumption of the network in WSNs.

As the number of sensor-enabled physical devices connected to the internet has dramatically increased, it is crucial for data to be transferred from source to destination as quickly as possible. So, routing is important in the Internet of Things. However, IoT is mobile by its very nature. Mobility is a good contender for effectively addressing hand-off time concerns, data transmission delays, overhead, and low packet delivery rates. The requirements for IoT routing protocols change constantly depending on the application. IoT have risen to the forefront of medical media technologies due to their small size and capacity for wireless data transport. High energy efficiency, transmission reliability, and extended battery life of sensor devices are necessary for a dependable network. The effectiveness of healthcare delivery is increased by taking protocol layers, data routing, and energy optimization measures into account. The need for steady, dependable, and real-time transmission due to the sizeable volume of data makes it imperative to find a solution. Numerous heterogeneous devices, a high bit error rate,

frequent network failures, and QoS assurance are the main problems with routing protocols. The quick and broad use of the Internet of Things (IoT) around the world has boosted the significant performance attained in terms of applications, technology, and security. Finally, we address the issues in IoT network, communication overhead, energy consumption and congestion issues for an optimal reliable solution. To overcome those problems in previous studies, COOC routing algorithm is proposed for IoT networks. The main objectives of COOC routing algorithm is describes as follows:

- Optimize clustering and multipath routing is used to formulate communication overhead aware optimal cluster-based (COOC) routing.
- By using COOC routing, we were able to increase network throughput and reduce computation and communication costs.
- Furthermore, COOC routing reduces the battery-power consumption of the network, increasing its overall lifetime.
- NS-2 simulator is used to evaluate our COOC routing.

#### B. Network Model of our COOC Routing Algorithm

Fig. 1 shows the typical structure of IoT network with our proposed COOC routing algorithm using optimal clustering and efficient multipath routing. The routing protocol is then used to send the data gathered from IoT sensors to the base station (BS). Almost all of the IoT network's gadgets run on limited, non-rechargeable energy sources like batteries. IoT applications typically operate in crowded, harsh locations, making it difficult to add or swap out the sensors' power sources.

The k-means clustering, fuzzy logic and genetic algorithm is used for the cluster formation using the basic information of IoT nodes location and distance between nodes to BS. Then, the rank of the nodes is compute by the individual position of nodes with respect to other nodes using multiple design constraints. The rank decreases in the up direction and increases in the down direction. Next, we develop an I-COOT algorithm for design constraints optimization which used to select CH among multiple nodes. Finally, we find the optimal best path between IoT nodes among multiple paths by using CGSO algorithm.

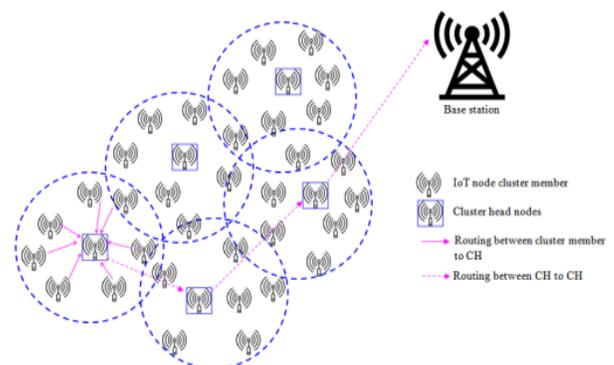


Fig. 1. Typical structure of IoT network with our COOC routing algorithm

#### IV. PROPOSED METHODOLOGY

We describe working function of our COOC routing algorithm which consists following set of process are clustering, design constraints for rank computation, optimization of design constraints and optimal path selection.

##### A. Clustering using Benchmark Algorithms

The creation of energy-efficient solutions becomes crucial since IoT nodes are energy-constrained and run on a small internal battery. In order to prepare for impending demand, energy conscious IoT networks must simultaneously forecast their energy use. A collection of sensor nodes that can sense, calculate, and transmit make up the network. Energy conservation in IoT becomes a major concern to increase network lifetime. Since clustering is regarded as an efficient and suitable way for transmitting the data without any issues, multiple efforts have been made to improve the routing protocols in the network to date. In this study, we used the k-means clustering, fuzzy logic, and genetic algorithms as three benchmark load-balanced clustering methods. It is necessary for this particular application to illustrate the chromosomal distribution to utilize the k-means clustering technique to partition unsatisfactory groupings. Our supposition is that the population is divided into clusters.

$$cq^{i,*} = \frac{1}{m_i} \sum_{c \in c_i} cq_N, \quad i=1,2,\dots,k \quad (1)$$

Where the number of cluster-related elements is  $m_i$ . The  $f$  symbol represents the altered feature space with a greater or even infinite dimension, and  $Y$  stands for the data space. The following objective function is minimized by KFCM.

$$I_{kfcM}(u, v) = \sum_{K=1}^C \sum_{j=1}^N \mu_{Kj}^M |\Phi(y_K) - \Phi(v_j)|^2 \quad (2)$$

The difference between the sizes of the largest cluster and the smallest cluster normalizes the size of the cluster and the size of the cluster.

$$\hat{H}_A = \frac{H_A - H_{\min}}{H_{\max} - H_{\min}} \quad (3)$$

$$\hat{H}_w = \frac{H_w - H_{\min}}{H_{\max} - H_{\min}} \quad (4)$$

Where  $H_{\max}$  and  $H_{\min}$  represent the normalized values, which range from zero to one, respectively. By explicitly detecting its presence in the algorithm, this circumstance can be avoided. Fuzzy logic provides the ability to make defensible conclusions in a world of uncertainty, imprecision, and missing data. It is therefore the optimal strategy to use in scenarios with real, continuous-valued elements because it uses data acquired in surroundings that include such qualities. The aforementioned context is appropriate for the information gathered about computer network traffic, which supports its use in anomaly detection. The membership degree of an element is obtained using a fuzzy membership function, which accepts a variety of arguments. The Gaussian membership function is an illustration of such a function.

$$\xi = E^{-\frac{(y-\hat{y})^2}{2\theta^2}} \quad (5)$$

Where  $\theta$  is a parameter that defines the standard deviation,  $\hat{y}$  is the center,  $y$  is the value to calculate its membership. In this paper, the function is derived from the Gaussian membership function as

$$\xi_k = 1 - E^{-\frac{(y_k - \hat{y}_k)^2}{2\theta_k^2}} \quad (6)$$

Using fuzzy logic, it is possible to determine whether an abnormality is happening right now. A fuzzy method is used to reduce this issue without impairing the system's capacity to detect anomalies.

##### B. Genetic Algorithm

Genetic Algorithm (GA) is a worldwide, parallel, stochastic search approach that exhibits significant robustness in problem domains where formal, strict, classical analysis is not feasible. The roulette wheel and a competition are two of these selection techniques. The odds of winning in roulette are determined by the fitness values of the chromosomes, which dictate

$$q_j = \frac{F_j}{\sum_{i=1}^N F_i} \quad (7)$$

Based on the results of the trials, it can be concluded that fuzzy logic and k-means clustering are not as exact as the best option for categorization. The precision and recall are both greatly enhanced by the GA algorithm. The definitions of recall and precision

$$precision = \frac{tp}{tp + fp} \quad (8)$$

$$recall = \frac{tp}{tp + fn} \quad (9)$$

The superiority of the GA algorithm becomes more apparent and we can get the meaningful findings more quickly when it is applied to a few additional data sets.

$$precision = \frac{|relevant \cap retrieved|}{|retrieved|} \quad (10)$$

$$recall = \frac{|relevant \cap retrieved|}{|testing|} \quad (11)$$

When the user is aware of the nonlinearities in the problem, the GA method can perform pretty well. However, the GAKFCM is more accurate and can overcome problems for the GA algorithm. And take a different look at the GA algorithm. The workings of cluster construction employing k-means clustering, fuzzy logic, and genetic algorithm are described in algorithm 4.1.

**Algorithm 4.1** Benchmark algorithm for cluster formation

Input: location, distance between nodes and BS

Output: Formation of cluster

1. Initiate the random population
2. Minimize the KFCM by objective function
 
$$I_{kfcM}(u, v) = \sum_{K=1}^C \sum_{j=1}^N \mu_{Kj}^M |\Phi(y_K) - \Phi(v_j)|^2$$
3. The condition is avoided by specifically checking for it within the algorithm
4.  $j=0$  and  $i=1$
5. Define Gaussian membership function
 
$$\xi = E \frac{-(y-\hat{y})^2}{2\theta^2}$$
6. Find fitness values of the chromo  $q_j = \frac{F_j}{\sum_{i=1}^N F_i}$
7. Get the useful results faster
 
$$precision = \frac{|relevant \cap retrieved|}{|retrieved|}$$

$$recall = \frac{|relevant \cap retrieved|}{|testing|}$$
8. Update the final values
9. End

C. Design Constraints Optimization

The practice of minimizing the amount of input constraints when creating a predictive model is known as design constraint optimization. In some circumstances, less input constraints might increase model performance while also lowering the computing cost of simulations. Using the Improved COOT bird optimal search method (I-COOT), this work chooses the cluster head (CH) based on the rank of each node. Coots are little waterfowl that belong to the Rallidae family of rails. They belong to the Fulica genus, which is named after the Latin word for "coot." This bird's actions on the water's surface can be used as an optimization technique. Coots appear to be well within what is, for surf scoters, a zone of repulsion as they travel at an angle to their direction of motion. There is no assurance that a solution will be found in one run when using population-based optimization approaches to discover the ideal number of optimization issues. However, if there are enough random solutions and optimization processes, the likelihood of discovering the overall best improves. Using the formula, the population is produced at random in the small area.

$$cootpos(j) = rand(1, D) * (ua - la) + la \quad (12)$$

where  $cootpos(j)$  is the coot position,  $d$  the number of variables or problem dimensions,  $la$  is the lower bound of the search space and ' $ua$ ' is the upper bound of the search space. Each variable may have a different lower bound and upper bound problem.

where  $D$  is the number of variables or problem dimensions, and  $lb$  and ' $ua$ ' represent the lower and upper bounds of the search space, respectively. Coot position is

$cootpos(j)$ , there could be many lower bound and upper bound issues for each variable.

$$la = [la_1, la_2, \dots, la_D], ua = [ua_1, ua_2, \dots, ua_D] \quad (13)$$

In order to carry out this movement, we take into account a random place within the search space and move the coot in that direction.

$$P = rand(1, D) * (ua - la) + la \quad (14)$$

The search space is explored by this coot movement in many areas. This movement will let the algorithm escape the local optimal if it becomes stuck in the local optimal. The new position of the coot is calculated as follows:

$$cootpos(j) = cootpos(j) + B * r2 * (P - cootpos(j)) \quad (15)$$

we compute  $B$  using the random movement of the coot in various directions, where " $r2$ " is a random value in the range  $[0, 1]$ .

$$B = 1 - l * \left( \frac{1}{Iter} \right) \quad (16)$$

Where,  $Iter$  is the maximum iteration and  $l$  is the current iteration. The typical alignments of two coots are used for implementing chain movement. We may also move the coot toward the other coot by roughly halving the distance between them after first calculating the distance vector between them. We employed the first technique, and a formula was applied to determine the coot's new position.

$$cootpos(j) = 0.5 * (cootpos(j - 1) + cootpos(j)) \quad (17)$$

The group is often led by a few coots in the front, and the remainder of the coots must move closer and alter their posture in accordance with the group's leaders. One possible query is if each coot will change its position according to which leader. The coots can adjust their position based on the average position of the leaders, which can be taken into consideration. Premature convergence results from taking the average position into account. We employ a system in accordance with the leader-selection process to carry out this movement.

$$k = 1 + (j \bmod nl) \quad (18)$$

where  $K$  is the leader's index number,  $NL$  is the total number of leaders, and  $I$  is the index number of the current coot. Based on the leader's  $k$ , the coot ( $j$ ) must update its location to determine the coot's subsequent position based on the chosen leader.

$$cootpos(j) = leaderpos(K) + 2 * r1 * \cos(2r\pi) * (leaderpos(K) - cootpos(j)) \quad (19)$$

Where, the coot's current position is  $cootpos(j)$ ,  $leaderpos(K)$  has been chosen as the leader position.  $R1$  is a random number between 0 and 1,  $\pi$  is the same as pi, or 3.14, and ' $r$ ' is a random number between 0 and 1. Around this current ideal location, this formula seeks out better positions. Leaders may need to shift away from the present best position in order to find better positions. This formula offers a useful method for moving away from and toward the ideal place.

$$leaderpos(j) = \begin{cases} A \times r3 \times \cos(2r\pi) \times & r4 < 0.5 \\ (gbest - leaderpos(j)) + gbest & \\ A \times r3 \times \cos(2r\pi) \times & r4 \geq 0.5 \\ (gbest - leaderpos(j)) - gbest & \end{cases} \quad (20)$$

Where the best position ever discovered is  $gbest$ , R3 and R4 are random numbers in the range [0, 1], is equal to 3.14 pi, R is a random number in the range [1, 1], and B is determined using the formula.

$$A = 2 - l \times \left( \frac{1}{Iter} \right) \quad (21)$$

Where 'Iter' for the most iterations and 'l' stands for the current iteration. 2 x r3 greater random movements are made to prevent the algorithm from becoming stuck at the local optimum. This indicates that while we are in the exploitation phase, we are simultaneously undertaking exploration. Cos(2rπ) looks for a better position near the best search agent by searching in various radii about it. Algorithm 4.2 describes the working process of CH selection using I-COOT.

**Algorithm 4.2** CH selection using I-COOT

Input : Energy efficiency, Link quality, Path loss, Distance and Delay  
Output : design constraints optimization and CH

- 1 Initialize the random population
- 2 The population is randomly generated  
 $cootpos(j) = rand(1, D) \cdot (ua - la) + la$
- 3 J=0, and i=1
- 4 Define coot towards this random position  
 $P = rand(1, D) \cdot (ua - la) + la$
- 5 Selection of leader  $k = 1 + (j \bmod nl)$
- 6 Selection of coot based on selected leader  
 $cootpos(j) = leaderpos(K) + 2 \times r1 \times \cos(2r\pi) \times (leaderpos(K) - cootpos(j))$
- 7 Compute B  $A = 2 - l \times \left( \frac{1}{Iter} \right)$
- 8 Update the final values
- 9 End

**D. Optimal Path Selection**

A well-known trade-off in the architecture of IoT is minimizing power consumption at the expense of the performance of the network. Traditional sensor network platforms were created with a focus on low power consumption at the expense of communication throughput. To collect auditory and visual data, which has a high need for transmission throughput, new apps are being used. According to the investigation, conventional ways fail to deliver improved security and quality of service (QoS), as well as to balance the temperature and load of WSN-IoT devices. They also fail to extend the network lifetime and reduce energy depletion. In order to ensure optimal data transfer from CHs, we have created the chaotic golden search optimization algorithm (CGSO), which selects the best path among numerous paths connecting IoT nodes. Kiefer introduced the golden section search (GSO) in 1953. (Kiefer, 1953). When an

object function is uni-modal, this approach can be used. The approach performs admirably when solving object functions that are either impossible to discriminate or difficult to differentiate. 2-D GSS for object tracking is a newly introduced variation of the golden section search. It also appears in straightforward maps like the logistic map. Typically, a one-dimensional chaotic map looks like this:

$$y(N + 1) = F(\mu_1, \mu_2, \dots, \mu_M, y(N)), N = 0, 1, 2, 3 \dots \quad (22)$$

A chaotic map is fused with GSO algorithm to optimize path selection constraints. The chosen control parameters  $\mu_j, j = 1, \dots, M$  are quite modest, yet even a little shift in the chaotic variable's starting value, x, will have a significant impact on subsequent values of the chaotic variable, y. To define one-dimensional chaotic maps as follows:

$$y(N + 1) = by(N)(1 - y(N)) y(0) \in (0, 1), \quad (23)$$

$$y(0) \notin \{0, 0.25, 0.5, 0.75, 1\}$$

$$y(N + 1) = \cos(K \cos^{-1}(y(N))) y \in (-1, 1) \quad (24)$$

Chaos with b=4 is generated by the logistic map. le=0.6932 is the Lyapunov exponent of the Chebyshev map with k=2. The scout bee uses abandoned food sources as new ones.

$$Y_{ji} = y_{min,i} + f(y_{max,i} - y_{min,i}) \quad (25)$$

Depending on the scale factor  $f_1$  &  $f_2$  the creation of new food sources is viewed as a black box procedure. The primary concept is that, in order to ensure a high-quality solution that will play a crucial part in succeeding generations, the updating of the scale factor and, consequently, the creation of the food sources, are, with a given probability, controlled. The procedure introduces the Chaotic Golden Search Optimization Technique (CGSO), a traditional local search algorithm for non-differentiable fitness functions. In order to produce high-quality food sources, the scale factor golden section search uses the golden section search to scale factor. This procedure produces two intermediate points in the range [a = 1, b = -1]:

$$f_1 = a - \frac{a-b}{\delta}, \quad (26)$$

$$f_2 = b + \frac{a-b}{\delta} \quad (27)$$

The scaling factor's upper and lower bound values are calculated as follows. In the GSS algorithm, the first two points  $y_1, y_2 \in [l, u]$  are calculated as follows.

$$C = \frac{-1 + \sqrt{5}}{2} \quad (28)$$

$$y_1 = Cl + (1 - C)u \quad (29)$$

$$y_2 = (1 - C)l + Cu \quad (30)$$

The search is carried out until the stop criteria are met. As a result, just one of these portions is chosen for the subsequent iteration. As a result, it is essential that the two parts are of equal width. In certain circumstances, the bigger portion is

therefore taken more repeatedly, and the convergence speed is slowed.

$$Q + P = P + R \quad (31)$$

$$\frac{Q}{P} = \frac{Q}{P} = \frac{(P+Q)}{Q} = \frac{1}{C} = \varphi \quad (32)$$

From above equations, we follow that  $\varphi=161803398\dots$  and  $C=1.61803398\dots$ . Thus, if  $n$  is the number of iterations then,  $\varphi^n$  is the convergence rate of CGSO algorithm. The search interval shrunk to less than 1.0% of the original interval for  $n=15$ . The algorithm 4.3 describes the working function of optimal path selection using CGSO. We follow from the aforementioned equations that  $\varphi=161803398\dots$  and  $C=1.61803398\dots$ . As a result, if  $n$  is the number of iterations, then the CGSO algorithm's rate of convergence is  $\varphi^n$ . Less than 1% of the initial search interval for  $n=15$  remained after the search interval shrank. The operation of the CGSO-based optimal path selection algorithm is described in algorithm 4.3.

**Algorithm 4.3** Chaotic Golden Search Optimization algorithm (CGSO)

Input: CH, congestion rate, aggregation delay

Output: optimal paths

- 1 Initialize the random population
- 2 Define one-dimensional chaotic map  
 $y(N+1) = F(\mu_1, \mu_2, \dots, \mu_M, y(N)), N = 0, 1, 2, 3, \dots$
- 3 The interval values are  $a = 1, b = -1$
- 4 Define the initial food sources  
 $Y_{ji} = y_{\min,i} + f(y_{\max,i} - y_{\min,i})$
- 5 Calculate two points in GSS algorithm:  
 $y_1, y_2 \in [l, u] \quad C = \frac{-1 + \sqrt{5}}{2}$
- 6 The convergence speed is reduced  $Q + P = P + R$
- 7 Update the final values
- 8 End

V. RESULTS AND DISCUSSION

We describe simulation results and comparative analysis of suggested COOC and existing routing algorithms with the different simulation scenarios. COOC routing algorithm is simulate and analyze using Network simulator (NS-2). The simulation results are done for performance evaluation of the COOC routing algorithm against existing state-of-art algorithms' performance, based on clustering FPT (CFPT) [28], routing FPT-approximation (RFPT) [29], energy-efficient and EB-CRP (energy-balanced cluster-based routing protocol) [30] and fixed-parameter tractable approximation clustering (FPTAC) [27].

A. Simulation Setup

The sink is situated in the centre of a terrain measuring 1000 m by 1000 m, where the nodes are situated. A grid of 100 nodes is created, and the remaining 900 are distributed at random. The sensor node's transmission and reception power consumption are 24.92 and 19.72 mJ per byte, respectively,

according to the simulation requirements. IoT sensors come in a variety of numbers, ranging from 200 to 1000. Each sensor node and each gateway are assumed to have a starting energy of 2J. Each node communicates with the others via the CSMA/CA MAC layer protocol. To send the data packets to the gateways, the sensor nodes employ the TDMA that the BSs choose. The IoT sensor nodes' transmission range is 100 metres, and the data packet size is 4000 bits. The simulations are run 30 times, and the accompanying graph shows the typical outcome of the runs. The precise configuration of our simulation is described in Table II.

TABLE II. SIMULATION SETUP

Simulation Area	1000m×1000m
Number of IoT sensors	200-1000
Data size	4000 bits
Control packet size	200 bits
Senor sensing range	80 m
Initial energy of sensor nodes	2J
MAC protocol	CSMA/CA
Bandwidth	250 Kb/s
Payload size	30 bytes
Transmission range	100 m
Avg. energy consumption of transmitting node	24.92mJ per byte
Avg. energy consumption of receiving node	19.72mJ per 1 byte
Simulation time	30 times

B. Comparative Analysis on Routing Algorithms

1) *Impact of node density:* Using 200, 400, 600, 800, and 1000 nodes as well as a fixed network size of 1000m×1000m, we analyze the performance of our proposed and existing routing algorithms. A simulation analysis of existing and proposed routing algorithms, energy consumption, throughput, network lifetime, routing overhead, reception ratio, and average link lifetime is presented. Energy consumption of our proposed and existing routing algorithms is compared in Table III and Fig. 2. By utilizing node density as an indicator of energy consumption performance, the proposed COOC routing algorithm ensures better solution. The energy consumption of our proposed COOC routing algorithm is 21.685%, 17.196%, 12.161% and 6.474% efficient than the existing CFPT [28], RFPT [29], EB-CRP [30], and FPTAC [27] routing algorithms respectively. The throughput of our proposed and existing routing algorithms is compared in Table IV and Fig. 3. It appears that the COOC scheme has the most throughput compared to the other routing schemes, with a value of 80300Mbps for 200 nodes and 66000Mbps for 1000 nodes. The CFPT scheme has the lowest throughput, with a value of 25000Mbps for 200 nodes and 8000Mbps for 1000 nodes. Overall, the results suggest that the COOC scheme provides the best performance in terms of throughput, followed by FPTAC, EB-CRP, RFPT and CFPT.

TABLE III. ENERGY COMPARISON (J) WITH NODE DENSITY

Routing scheme	Number of nodes				
	200	400	600	800	1000
CFPT	173.48	274.4	558.76	958.34	1229.3
RFPT	137.56	236.41	522.22	919.08	1195.3
EB-CRP	97.28	197.20	485.73	887.08	1139.3
FPTAC	51.79	159.52	449.71	850.52	1106.5
COOC	12.53	121.51	412.5	812.54	1070.5

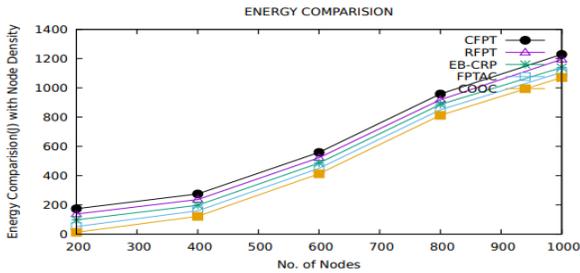


Fig. 2. Energy consumption with node density

TABLE IV. THROUGHPUT WITH NODE DENSITY

Routing scheme	Number of nodes				
	200	400	600	800	1000
CFPT	25000	20000	18080	11000	8000
RFPT	40000	35008	30000	27500	22391
EB-CRP	54000	50000	45000	40100	37000
FPTAC	69000	65000	60000	57000	51000
COOC	80300	79900	73123	68500	66000

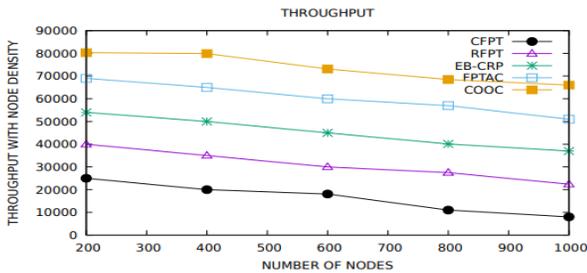


Fig. 3. Throughput with node density

The network lifetime of our proposed and existing routing algorithms is compared in Table V and Fig. 4. By utilizing node density as an indicator of network lifetime performance, the proposed COOC routing algorithm ensures better solution. The network lifetime of our proposed COOC routing algorithm is 26.147%, 19.09%, 13.073% and 6.537% efficient than the existing CFPT [28], RFPT [29], EB-CRP [30], and FPTAC [27] routing algorithms respectively. The routing overhead of our proposed and existing routing algorithms is compared in Table VI and Fig. 5. By utilizing node density as an indicator of routing overhead performance, the proposed COOC routing algorithm ensures better solution.

TABLE V. NETWORK LIFETIME WITH NODE DENSITY

Routing scheme	Number of nodes				
	200	400	600	800	1000
CFPT	28.02	27.05	26.16	24.27	23.89
RFPT	33.07	30.12	28.46	27.49	26.04
EB-CRP	34.02	33.25	30.30	28.16	27.01
FPTAC	35.05	34.34	33.32	32.43	30.35
COOC	38.70	36.14	35.05	34.41	33.17

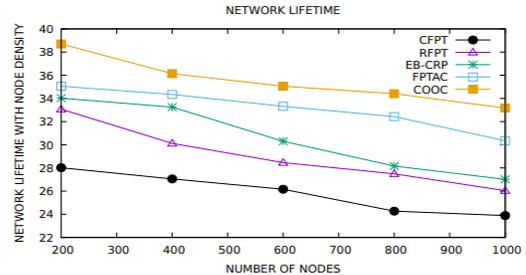


Fig. 4. Network lifetime with node density

TABLE VI. ROUTING OVERHEAD WITH NODE DENSITY

Routing scheme	Number of nodes				
	200	400	600	800	1000
CFPT	65.01	69.15	73.21	77.21	80.12
RFPT	63.24	67.61	70.16	74.92	78.73
EB-CRP	60.12	63.97	67.24	72.84	75.12
FPTAC	57.86	62.62	65.13	68.71	73.34
COOC	55.12	59.21	63.89	67.29	70.10

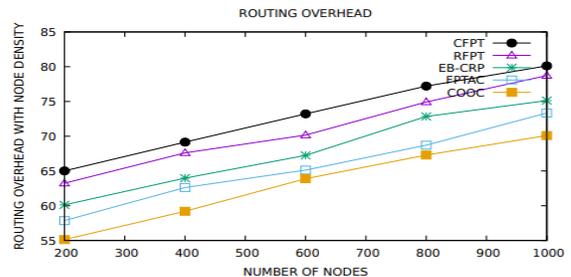


Fig. 5. Routing overhead with node density

TABLE VII. RECEPTION RATIO WITH NODE DENSITY

Routing scheme	Number of nodes				
	200	400	600	800	1000
CFPT	64.02	63.12	61.42	60.12	59.29
RFPT	72.12	70.44	68.12	65.69	64.90
EB-CRP	82.24	80.15	78.60	76.75	75.50
FPTAC	90.21	89.21	88.12	87.30	86.21
COOC	98.13	97.41	96.3	95.20	94.12

The routing overhead of COOC routing algorithm is 13.465%, 10.451%, 7.219%, and 3.744% efficient than the existing CFPT [28], RFPT [29], EB-CRP [30], and FPTAC [27] routing algorithms respectively.

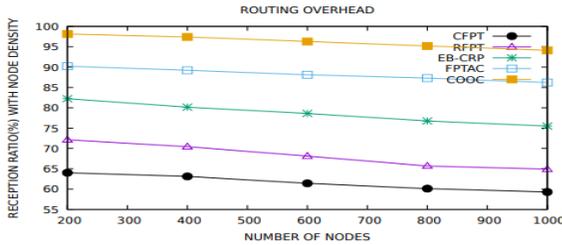


Fig. 6. Reception ratio with node density

TABLE VIII. AVERAGE LINK LIFETIME WITH NODE DENSITY

Routing scheme	Number of nodes				
	200	400	600	800	1000
CFPT	12.51	11.25	10.14	8.75	7.51
RFPT	15.21	13.75	12.52	11.25	10.54
EB-CRP	17.25	16.16	14.57	13.48	11.75
FPTAC	19.53	18.25	17.54	15.75	14.51
COOC	22.01	20.75	19.51	18.25	17.48

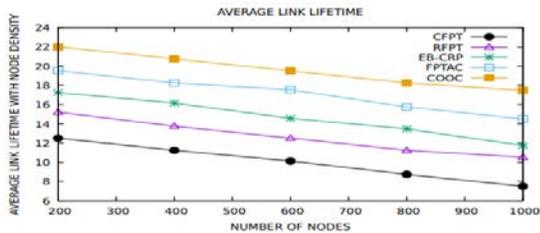


Fig. 7. Average link lifetime with node density

The reception ratio of proposed and existing routing algorithms is compared in Table VII and Fig. 6. By utilizing node density as an indicator of reception ratio performance, the proposed COOC routing algorithm ensures better solution. The reception ratio of COOC routing algorithm is 32.119%, 21.94%, 16.059% and 8.03% efficient than the existing CFPT [28], RFPT [29], EB-CRP [30], and FPTAC [27] routing algorithms respectively. The average link lifetime of proposed and existing routing algorithms is compared in Table VIII and Fig. 7. It appears that the COOC scheme has the most average link lifetime compared to the other routing schemes, with a value of 22.01 seconds for 200 nodes and 17.48 seconds for 1000 nodes. The CFPT scheme has the lowest average link lifetime, with a value of 12.51 seconds for 200 nodes and 7.51 seconds for 1000 nodes. Overall, the results suggest that the COOC scheme provides the best performance in terms of average link lifetime, followed by FPTAC, EB-CRP, RFPT and CFPT.

## VI. CONCLUSION

In this work, based on hybrid heuristic, we propose communication overhead aware optimal cluster-based (COOC) routing algorithm for IoT networks. With the use of

k-means clustering, fuzzy logic, and genetic algorithms, we form load-balanced clusters. I-COOT is used to optimize multiple design constraints to compute the rank of each node in a cluster. In IoT networks, we reduce communication overhead by select CH according to the rank condition. A chaotic golden search optimization algorithm (CGSO) is designed for optimizing data transfer from the CHs by identifying the best path among multiple paths among IoT nodes. In conclusion, we validate our proposed COOC routing algorithm against different simulation scenarios. From the simulation results, we observed that the effectiveness of our proposed COOC routing algorithm perform very effective manner in terms of consumption, throughput, network lifetime, routing overhead, reception ratio, and average link lifetime compared to existing routing algorithms.

## REFERENCES

- [1] Li, X., Qi, H. and Wu, J., 2022. Node social nature detection OSN routing scheme based on IoT system. *IEEE Internet of Things Journal*.
- [2] Ghosh, S., Dagiuklas, T., Iqbal, M. and Wang, X., 2022. A cognitive routing framework for reliable communication in iot for industry 5.0. *IEEE Transactions on Industrial Informatics*, 18(8), pp.5446-5457.
- [3] Deng, S., Zhao, H., Xiang, Z., Zhang, C., Jiang, R., Li, Y., Yin, J., Dustdar, S. and Zomaya, A.Y., 2021. Dependent Function Embedding for Distributed Serverless Edge Computing. *IEEE Transactions on Parallel and Distributed Systems*, 33(10), pp.2346-2357.
- [4] Yu, C., Shen, S., Yang, H., Zhang, K. and Zhao, H., 2021. Leveraging Energy, Latency and Robustness for Routing Path Selection in Internet of Battlefield Things. *IEEE Internet of Things Journal*.
- [5] Ghahramani, M., Zhou, M., Molter, A. and Pilla, F., 2021. IoT-based Route Recommendation for an Intelligent Waste Management System. *IEEE Internet of Things Journal*.
- [6] Wang, X., Hu, J., Lin, H., Garg, S., Kaddoum, G., Jalilpiran, M. and Hossain, M.S., 2022. QoS and privacy-aware routing for 5G enabled industrial internet of things: A federated reinforcement learning approach. *IEEE Transactions on Industrial Informatics*.
- [7] Zhang, Y., Ren, Q., Song, K., Liu, Y., Zhang, T. and Qian, Y., 2021. An Energy Efficient Multi-Level Secure Routing Protocol in IoT Networks. *IEEE Internet of Things Journal*.
- [8] Ma, N., Zhang, H., Hu, H. and Qin, Y., 2021. ESCVAD: An Energy-Saving Routing Protocol Based on Voronoi Adaptive Clustering for Wireless Sensor Networks. *IEEE Internet of Things Journal*, 9(11), pp.9071-9085.
- [9] He, Y., Han, G., Xu, M. and Martínez-García, M., 2021. A Pseudo-Packet Scheduling Algorithm for Protecting Source Location Privacy in the Internet of Things. *IEEE Internet of Things Journal*.
- [10] Liang, J., Liu, W., Xiong, N.N., Liu, A. and Zhang, S., 2021. An intelligent and trust UAV-assisted code dissemination 5G system for industrial Internet-of-Things. *IEEE Transactions on Industrial Informatics*, 18(4), pp.2877-2889.
- [11] H. Bany Salameh, R. Derbas, M. Aloqaily and A. Boukerche, "Secure Routing in Multi-hop IoT-based Cognitive Radio Networks under Jamming Attacks", *Proceedings of the 22nd International ACM Conference on Modeling, Analysis and Simulation of Wireless and Mobile Systems - MSWIM '19*, 2019. Available: 10.1145/3345768.3355944.
- [12] H. Rhim, R. Abassi, K. Tamine, D. Sauveron and S. Guemara, "A secure network coding-enabled approach for a confidential cluster-based routing in wireless sensor networks", *Proceedings of the 35th Annual ACM Symposium on Applied Computing*, 2020. Available: 10.1145/3341105.3373969.
- [13] T. Mick, R. Tourani and S. Misra, "LASeR: Lightweight Authentication and Secured Routing for NDN IoT in Smart Cities", *IEEE Internet of Things Journal*, vol. 5, no. 2, pp. 755-764, 2018. Available: 10.1109/jiot.2017.2725238.

- [14] G. Hatzivasilis, I. Papaefstathiou and C. Manifavas, "SCOTRES: Secure Routing for IoT and CPS", IEEE Internet of Things Journal, vol. 4, no. 6, pp. 2129-2141, 2017. Available: 10.1109/jiot.2017.2752801.
- [15] K. Haseeb, N. Islam, A. Almogren, I. Ud Din, H. Almajed and N. Guizani, "Secret Sharing-Based Energy-Aware and Multi-Hop Routing Protocol for IoT Based WSNs", IEEE Access, vol. 7, pp. 79980-79988, 2019. Available:10.1109/access.2019.2922971.
- [16] K. Haseeb, N. Islam, A. Almogren, and I. Ud Din, "Intrusion Prevention Framework for secure routing in WSN-based mobile internet of things," IEEE Access, vol. 7, pp. 185496–185505, 2019.
- [17] J. Duan, D. Gao, D. Yang, C. H. Foh, and H.-H. Chen, "An energy-aware trust derivation scheme with game theoretic approach in wireless sensor networks for IoT applications," IEEE Internet Things J., vol. 1, no. 1, pp. 58–69, 2014.
- [18] F. Wu et al., "An efficient authentication and key agreement scheme for multi-gateway wireless sensor networks in IoT deployment," J. Netw. Comput. Appl., vol. 89, pp. 72–85, 2017.
- [19] A. Onasanya, S. Lakkis, and M. Elshakankiri, "Implementing IoT/WSN based smart Saskatchewan Healthcare System," Wirel. netw., vol. 25, no. 7, pp. 3999–4020, 2019.
- [20] J. K. Jain, "Secure and energy-efficient route adjustment model for internet of things," Wirel. Pers. Commun., vol. 108, no. 1, pp. 633–657, 2019.
- [21] S. A. Jesudurai and A. Senthilkumar, "An improved energy efficient cluster head selection protocol using the double cluster heads and data fusion methods for IoT applications," Cogn. Syst. Res., vol. 57, pp. 101–106, 2019.
- [22] M. López, A. Peinado, and A. Ortiz, "An extensive validation of a SIR epidemic model to study the propagation of jamming attacks against IoT wireless networks," Comput. netw., vol. 165, no. 106945, p. 106945, 2019.
- [23] V. Kumar, G. Sakya, and C. Shankar, "WSN and IoT based smart city model using the MQTT protocol," J. Discrete Math. Sci. Cryptogr., vol. 22, no. 8, pp. 1423–1434, 2019.
- [24] Han, T., Zhang, L., Pirbhulal, S., Wu, W. and de Albuquerque, V.H.C., 2019. A novel cluster head selection technique for edge-computing based IoMT systems. Computer Networks, 158, pp.114-122.
- [25] El-shafeiy, E., Sallam, K.M., Chakraborty, R.K. and Abohany, A.A., 2021. A clustering based Swarm Intelligence optimization technique for the Internet of Medical Things. Expert Systems with Applications, 173, p.114648.
- [26] Ahmed, G., Mehmood, D., Shahzad, K. and Malick, R.A.S., 2021. An efficient routing protocol for internet of medical things focusing hot spot node problem. International Journal of Distributed Sensor Networks, 17(2), p.1550147721991706.
- [27] Yarinezhad, R. and Sabaei, M., 2021. An optimal cluster-based routing algorithm for lifetime maximization of Internet of Things. Journal of Parallel and Distributed Computing, 156, pp.7-24.
- [28] R. Yarinezhad, S.N. Hashemi, A routing algorithm for wireless sensor networks based on clustering and an fpt-approximation algorithm, J. Syst. Softw. 155 (2019) 145–161.
- [29] R. Yarinezhad, S.N. Hashemi, Solving the load balanced clustering and routing problems in WSNs with an fpt-approximation algorithm and a grid structure, Pervasive Mob. Comput. 58 (2019) 101033.
- [30] R. Yarinezhad, S.N. Hashemi, Increasing the lifetime of sensor networks by a data dissemination model based on a new approximation algorithm, Ad Hoc Netw. 100 (2020) 102084.

# Supply Chain Network Model using Multi-Agent Reinforcement Learning for COVID-19

Tomohito Okada<sup>1</sup>, Hiroshi Sato<sup>2</sup>, Masao Kubo<sup>3</sup>

Test and Evaluation Command, Japan Ground Self Defense Force, Gotemba City Sizuoka Prefecture, Japan<sup>1</sup>  
Department of Computer Science, National Defense Academy in Japan, Yokosuka City Kanagawa Prefecture, Japan<sup>2,3</sup>

**Abstract**—The COVID-19 vaccination management in Japan has revealed many problems. The number of vaccines available was clearly less than the number of people who wanted to be vaccinated. Initially, the system was managed by making reservations with age group utilizing vaccination coupons. After the second round of vaccinations, only appointments for vaccination dates were coordinated and vaccination sites were set up in Shibuya Ward where the vaccine could be taken freely. Under a shortage of vaccine supply, the inability to make appointments arose from a failure to properly estimate demand. In addition, the vaccine expired due to inadequate inventory management, resulting in the vaccine being discarded. This is considered to be a supply chain problem in which appropriate supply could not be provided in response to demand. In response to this problem, this paper examines whether it is possible to avoid shortage and stock discards by a decentralized management system for easy on-site inventory control instead of a centralized management system in real world. Based on a multi-agent model, a model was created to redistribute inventory to clients by predicting future shortage based on demand fluctuations and past inventory levels. The model was constructed by adopting the Kanto region. The validation results of the model showed that the number of discards was reduced by about 70% and out-of-stocks by about 12% as a result of learning the dispersion management and out-of-stock forecasting.

**Keywords**—Supply chain management; agent based model; multi-agent reinforcement learning; COVID-19 vaccination

## I. INTRODUCTION

The vaccination with the COVID-19 virus vaccine for the pandemic is managed through a vaccination ticket and vaccination reservation system, and priority vaccination is given based on the risk of serious illness and the security of the medical care system, because the amount of vaccine that can be secured is limited and its supply is expected to be sequential. The ministry of health, labour and welfare also prioritizes vaccinations based on the risk of severe cases of the disease and the availability of healthcare [1]. In July 2021, the amount of vaccine supplied by the national government to local governments became significantly insufficient. This caused some local governments to temporarily suspend vaccination appointments and medical institutions that have been forced to reduce their supply of vaccine are forced to coordinate with applicants who have made reservations to postpone their vaccinations [2]. However, about 2.2 million doses of vaccine have been discarded due to expiration [3]. In light of the above, some areas disposed vaccines due to inadequate inventory management while some areas are

experiencing shortages. The essence of these problems is that inventory management and demand forecasting were not properly carried out. The only centralized control by the government or relevant ministries is not sufficient to manage the situation.

Thus, when the issue of vaccine shortage and disposal was widely recognized in the press, etc., the measures to be taken focused on the storage method from the pharmaceutical knowledge of vaccines and the proposed solution regarding the vaccination system, and the Ministry of Health, Labor and Welfare provided an explanation on how to ensure the vaccination system [4]. However, very few have addressed issues related to supply chain management regarding the increased demand for vaccines related to the growing number of COVID-19 virus cases and inventory management related to this demand.

This paper creates two model of inventory management and shipping plan. These are the centralized management model and the decentralized management model. The centralized management model ships vaccines based on demand from each municipality. The decentralized management model is a model in which each municipality uses reinforcement learning to manage inventory and forecast demand [5]. In this model, a vaccination site with sufficient inventory provides inventory to a vaccination site with insufficient inventory. The model is designed to verify how effectively vaccines can be utilized when each municipality takes the initiative in inventory management and shipping.

## II. PRIOR RESEARCH

Supply chain research has been conducted using various approaches to achieve various objectives, such as avoiding shortage, reducing excess inventory, and reducing costs including design and model proposals based on engineering knowledge [6], mathematical optimization of risk management methods [7], and realistic simulation models focusing on lead time [8]. These approaches have achieved some goals.

In recent years, machine learning, especially AI, has been widely utilized as a problem-solving method. For example, it is researched to assign various planning tasks to machine learning in the design (long-term strategy), planning (medium-term and short-term strategy), and execution (operational level) stages for proactive supply chain problem-solving approaches [9]; and it is researched to seek to make strategic decisions based on machine learning forecasts of

environmental changes such as demand fluctuations for passive factors [10].

This paper examines the effectiveness of vaccine inventory exchanges from a supply chain management perspective by reinforcement learning about changes in vaccine demand at vaccination sites.

### III. COMPOSITION OF THE SIMULATION MODEL

#### A. Building the SCM Model in MAS

There are so many players to study for the Supply Chain Management (SCM). These players generally include from producers to retailers etc. Each player collects information for the sales strategies to be made by management of service, inventory and cost. They which are under the control of upper headquarter office, make their own decisions within their responsibility [11] [12].

Multi-Agent Simulation (MAS) discusses the coordination of behavior in a set of autonomous intelligent agents [13]. This simulation can lead to the whole optimum of the collection in which each player decide to act on its own.

The characteristics of the vaccine supply chain are close to those of the Multi-Agent Simulation. Supply chain model using multi-agent system makes it possible to analyze what kind of supply chain management is appropriate.

The agent based model of supply chain management in this study is built by artisoc3.0. This software is based on java and is specialized for multi-agent simulation [14].

By using Multi-agent Simulation to study supply chain management, we can analyze the linkages between each agent [15], [16], [17], between clusters [18], resilience [19], [20], dynamic network model [21], [22] and etc.

#### B. The Way to Apply Reinforcement Learning

In this study, the concept of Q-learning was applied to COVID-19 vaccine inventory management in each agent. There are some studies that using Q-Learning to manage inventory with expiration dates [23][24][25]. The agents in a competitive supply chain take their decisions individually in a distributed environment and independent of one another. At the same time, they must coordinate their actions [26][27]. In this time, supply chain management is needed to the balance the decision between centralized management and decentralized management.

The state as Q-learning is to avoid out of stock. In supply chain management, there is a value called safety stock quantity, which is the minimum amount of inventory that should be maintained to avoid out of stock. The state is defined in which the inventory quantity always exceeds the safety stock quantity.

The action as Q-learning is the selection of suppliers and the amount of order. Each agent has two way to select supplier. One is the order to an upper supplier with regular and limited quantities. The other is the request surrounding vaccination sites to provide vaccine inventory if they have a surplus. Acquiring vaccine inventory from upper supplier takes time, but the agent will certainly have amount of vaccine inventory.

The orders between same agents are not sure if they are in vaccine inventory, but if they have a surplus, they can get inventory immediately.

The reward as Q-learning is the amount of inventory in excess of the safety stock quantity. If the agent collects more inventory, some vaccine inventory might expire and be discarded. It would also unnecessarily increase transportation requirements by shipping to other agents. For these reasons, it is important to maintain an appropriate amount of inventory

#### C. Basic Structure of the Agent Model

There are three types of supply chain agents to be constructed in this study: government agents, local government agents, and vaccination site agents. Each of these agents behaves autonomously and has the ability to collect information, process information, make decisions, and act on its own.

The government agent can ensure the stock of vaccines on a regular basis. Based on the amount of vaccines demanded by the local government agents, the government agent ships vaccines to the local government agents. In this case, priority is given to the areas with large demand.

The local government agent ships vaccine stocks based on the quantity requested by the vaccination site agents. In this case, priority is given to the locations with the largest demand.

The vaccination sites agents consume vaccine inventory by administering vaccinations. The number of vaccinations (demand) over the past 100 days is recorded, and the amount of vaccine requested from the municipal agents is calculated based on consumption fluctuations.

As for the deadline for vaccine consumption, it is assumed to be 40 days after the government agent secures the vaccine.

This simulation is performed as one step per day, and the simulation is performed for five-year periods.

The relationship between each agent is shown in the Fig. 1

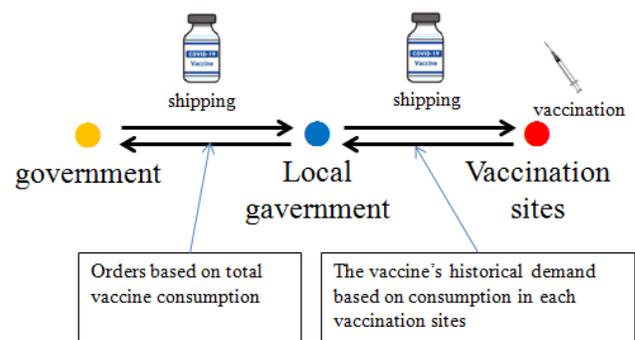


Fig. 1. The relationship between each agent

#### D. Centralized Management Model

1) Demand and supply: Each vaccination site should randomly vaccinate 0-50 persons per day. The reason for having variation due to randomness is that there are cases where the number of people who wish to be vaccinated continues to reach the daily limit in cases where vaccinations are given by appointment through the reservation system. In

addition, some vaccination sites that did not implement the reservation system had far more applicants than initially expected, resulting in shortage [28]. Conversely, there are days when the number of applicants for vaccination does not reach the allowable daily dose. Not only do such fluctuations in demand exist, but also there is always a steady-state shortage of vaccine throughout the country. To account for these fluctuations in demand and shortage conditions, the total number of vaccine stocks per 30 days was set in the simulation to 100,000 per 30 days, while the total demand is set to exceed this number. Since the total number of vaccination sites is 203, the average number of vaccine demand over the 30-day period is 100. The average number of vaccines in demand over a 30-day period is  $203 \times 30 \text{ days} \times 2.5 \text{ doses} = 152250 \text{ doses}$ .

2) *Vaccine supply method*: Local governments distribute vaccines based on demand at vaccination sites. In this case, the number of vaccines to be distributed is stated by the following formula. The lead time is 1 day.

$$\text{amount of shipping to vaccination sites} = \frac{\text{stock in local government} \times \text{one vaccination sites}}{\text{sum of demand in all vaccination sites}} \quad (1)$$

The amount of vaccine shipped by the government to local governments is based on the number of vaccination sites under the local government. This is stated by the following formula.

$$\text{amount of shipping to each local government} = \frac{\text{government stock} \times \text{number of vaccination sites in local government}}{\text{all vaccination sites}} \quad (2)$$

Suppose a country can obtain 100,000 doses of vaccine every 30 days.

**E. Decentralized Management Model**

1) *Differences from the centralized management model*: This model is used for the same model, as the centralized model of the demand specification and vaccine supply methods. The difference is that each vaccination sites has ability to receive the vaccine from other sites. When a vaccination center estimated the possibility of shortage of vaccine, it would inquire at other vaccination sites in the order of near to itself to see if there was any sufficient stock. If the other vaccination sites determine that they afford to tolerate sharing the vaccine stock, the vaccine can be shipped to other vaccination sites. In this case, the vaccination sites themselves calculate and set their own order time and order quantity based on the results of the reinforcement learning of fluctuations in vaccination demand and stock expiration dates.

2) *Reinforcement learning model*: The demand for vaccines and vaccine expiration dates for the past 100 days are recorded. This data is used to estimate future inventory status to prevent vaccine shortage through regression analysis. The objective variable was the amount of orders placed to other vaccination sites, and the explanatory variables were own demand and the number of days remaining before the

expiration date of the vaccine in their possession. An order quantity formula and conditions for order time are as follows.

$$\text{amount of order} = \text{average in demand} \times (\text{number of days left until the vaccine use deadline} + \text{lead time}) \quad (3)$$

$$\text{condition for order: stock} < (\text{number of days left until the vaccine use deadline} \times \text{average of demand}) \quad (4)$$

**IV. APPLICATION OF REAL DATA TO THE SIMULATION MODEL**

This simulation model is modeled after cities in the Kanto region. The government agent is a single agent modeled as the Ministry of Health, Labor and Welfare, which manages the importation of vaccines. The local government agents are seven agents modeled as Tokyo, Kanagawa, Saitama, Chiba, Gunma, Ibaraki, and Tochigi prefectures, which distribute vaccines to each vaccination site. A total of 203 vaccination site agents, modeled to designated cities, cities, and special wards, administer vaccines. The basic relationship between agents is shown in Fig. 2

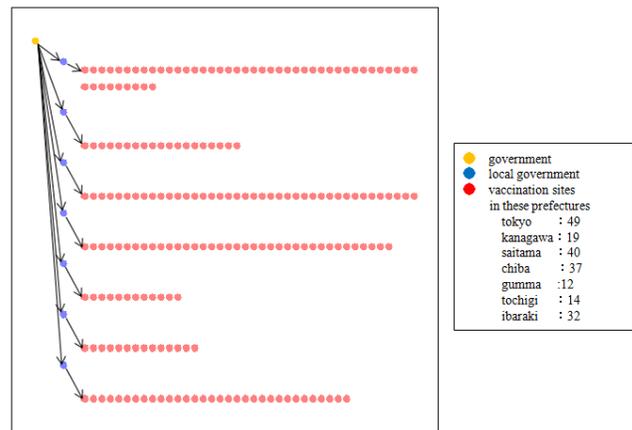


Fig. 2. Relationship diagram of each agent

**V. RESULTS**

**A. Results**

The results of the modeling and simulation according to the above are given in Table I.

TABLE I. SIMULATION RESULTS

Simulation No.	Centralized management model		Decentralized management model	
	Discarded number	Vaccine shortage	Discarded number	Vaccine shortage
1	82	28878	41	5456
2	64	28338	57	5206
3	72	27263	62	5389
4	73	28198	49	5402
5	84	27947	87	5411
average	75	28125	59	5373

Each simulation number is a five-year simulation, and "average in Simulation No." is the average of each experiment. Discarded number is the number of discarded vaccine. Vaccine shortage is the total number of vaccine demand in the absence of vaccine when there was a demand for vaccination.

### B. Considerations

Regarding the number of vaccine discards, the decentralized management model has reduced the number of discards to about 70% of the centralized management model. Shortage were also reduced to about 12% of the centralized management model. These indicate that more effective vaccination is possible when there is an exchange of vaccines among vaccination sites.

As for the reason for the number of discards, it was observed that when periods of extremely low demand occur consecutively, even vaccination sites with stock shortages are fully stocked, resulting in sufficient discarded vaccine. It was also observed that when demand increased during the above-mentioned period of reduced demand and small inventory, there was an overall shortage of vaccine to meet the demand, resulting in shortage.

## VI. CONCLUSION

In this paper, vaccine inventory management and shipping plan were simulated using a centralized management model and a decentralized management model to reveal the management of the vaccine demand and to avoid shortage. Compared to the model with centralized management, which caused problems in reality, the decentralized management model verified in the model reduced shortage of expired vaccines by approximately 70% and vaccine shortage by approximately 12%. It is estimated that the ability to exchange vaccines in the vicinity of vaccination sites where vaccines are consumed will greatly reduce the number of discards and the possibility of non-vaccination due to vaccine shortages.

## VII. FUTURE RESEARCH ISSUES

The next challenge is to consider supply chain management that considers multiple product elements and can streamline those that include more variables.

In Japan today, prices for food and many other things are rising. And it is said that security is needed for many things such as rare metals, semiconductors, oil, wheat, etc. But these problems are treated as vertical issues such as price increases for raw materials, processing cost, and so on. Many factors must be considered across the board to solve essential problems that affect the final product.

## ACKNOWLEDGMENT

The author appreciates Associate Professor Hiroshi Sato and Associate Professor Masao Kubo in National Defense Academy in Japan for providing help and instructions for this study.

## REFERENCES

[1] Ministry of Health, Labour and Welfare, "COVID-19 vaccine Q&A," <https://www.cov19-vaccine.mhlw.go.jp/qa/0047.html>, October 2022.

[2] Kenyu Sumie, "Urgent request for resolution of the new COVID-19 vaccine shortage and improvement of the immunization system," [https://hodanren.doc-net.or.jp/news/teigen/210706\\_yosei\\_cvd.pdf](https://hodanren.doc-net.or.jp/news/teigen/210706_yosei_cvd.pdf), 6th July 2021.

[3] The Asahi Shimbun, "Conventional vaccines to be disposed of in large quantities due to expiration. 2.2 million doses in stock in government ordinance cities", <https://www.asahi.com/articles/ASQB54G46QB3UTIL029.html>, October 2022.

[4] Ministry of Health, Labour and Welfare, "Securing the vaccination system for the new corona virus vaccine," <https://www.mhlw.go.jp/content/10906000/000708055.pdf>, Nov 2022.

[5] Tatsushi Nishi, "Distributed Optimization Technique for Supply Chain Management," Journal of the Japanese Society for Artificial Intelligence, Vol.19, No.5, pp571-578, September 2004

[6] Mikio Kubo, "Logistics engineering," Asakura publishing, 2001.

[7] Kazuhiro Kobayashi, "Mathematical Optimization techniques in transportation system management," Papers of national maritime research institute, vol.14, No.4, pp.303-320, 2014.

[8] Kengo Kami, "Supply chain model analysis under emergency conditions," Shodai Business Review in University of Hyogo, Vol.1, No.2, pp.17-30, 2012.

[9] Hannah Wenzel et al, "A literature review on machine learning in supply chain management," Artificial Intelligence and Digital Transformation in Supply Chain Management, pp.413-441, September 2019.

[10] Pooja Sareen, "Machine Learning and Supply Chain Management -A conceptual view," IJIRT, Vol.9, Issue.1, June 2022

[11] Nguyen Quoc Viet, Behzad Behdani, Jacqueline Bloemhof, "The value of information in supply chain decisions: A review of the literature and research agenda," Computers & Industrial Engineering, Vol.120, pp68-82, June 2018

[12] George Q. Huang, Jason S. K. Lau, K. L. Mak, "The impacts of sharing production information on supply chain dynamics: A review of the literature," International Journal of Production Research, Vol.41, issue 7, pp1483-1517, November 2010

[13] Yasuhiko Kitamura, "Multiagent," Knowledge base in The Institute of Electronics, Information and Communication Engineers of JAPAN, Ver.1, Section7-7-1, No.1-7, pp.17, 2019.

[14] Susumu Yamakage, "Instruction for building artificial society," shosekikoubouhayama, 2007

[15] B. Behdani et al, "Agent-based modeling to support operations management in a multi-plant enterprise," International Conference on Networking, Sensing and Control, pp323-328, May 2009

[16] Gonzalo Guillén et al, "An agent-based approach for supply chain retrofitting under uncertainty," Computer Aided Chemical Engineering, Vol.20, pp1555-1560, 2005

[17] Tomohito Okada, Akira Namatame, Hiroshi Sato, "An Agent-Based Model of Smart Supply Chain networks," Proceedings in Adaptation Learning and Optimization, Vol.5, pp373-384, November 2015

[18] Tomohito Okada, Akira Namatame, Hiroshi Sato, Saori Iwanaga "A Method to Reduce the Amount of Inventoried Stock in Thai Supply Chain," INTELLIGENT AND EVOLUTIONARY SYSTEMS, Vol.8, pp347-359, 2017

[19] Tomomi Kitou, "Structural Robustness of Real-world Supply Chains: A complex Network Approach," Information Processing Society of Japan. Transactions on mathematical modeling and its applications, Vol.6, No.2, pp174-181, August 2013

[20] Emma Brandon-jones, Brian Squire, Chad W. Autry, Kenneth J. Petersen, "A Contingent Resource-Based Perspective of Supply Chain Resilience and robustness," Journal of Supply Chain Management, Vol.50, Issue3, pp3-5, April 2014

[21] Jayashankar M. Swaminathan et al, "Modeling Supply Chain Dynamics: A Multiagent Approach," DECISION SCIENCES, Vol.29, Issue.3, pp607-632, June 2007

[22] D. R. Towill, M. M. Naim, J. Wikner, "Industrial Dynamics Simulation Models in the Design of Supply Chains," International Journal of Physical Distribution & Logistics Management, Vol.22, No.5, pp3-13, 1992

- [23] Guilherme O Ferreira, Edilson F Arruda, Lino G Marujo, "Inventory management of perishable items in long-term humanitarian operations using markov decision processes," *International Journal of Disaster Risk Reductions*, Vol.31, pp460-469, 2018
- [24] Ahmet Kara, Ibrahim Dogan, "Reinforcement Learning approaches for specifying ordering policies of perishable inventory systems," *Expert Systems with Applications*, Vol.91, pp150-158, January 2018
- [25] Yuuto Takahashi, Mitsuhiro Hoshino, "On an optimization method for perishable inventory problem using reinforcement learning," *Kyoto University Research Information Repository*, Vol.2214, pp27-38, April 2022
- [26] Tim van Tongeren et al, "Q-learning in a competitive supply chain," In converece proceedings - *IEEE International Conference on Systems, Man and Cybernetics*, pp1211-1216, January 2008
- [27] Y. Zhang, S.Bhattacharyya, "Effectiveness of Q-learning as a tool for calibrating agent-based supply network models," *Enterprise Information Systems*, Vol.1, pp217-233, May 2007
- [28] Tokyo Metropolitan Government, "Announcement from Tokyo Metropolitan Government Vaccine Large-Scale Vaccination Site," [https://twitter.com/tocho\\_vaccine/status/1431920909267308545](https://twitter.com/tocho_vaccine/status/1431920909267308545), August 2021.

# An Enhanced MCDM Model for Cloud Service Provider Selection

Ayman S. Abdelaziz<sup>1\*</sup>, Hany Harb<sup>2</sup>, Alaa Zaghoul<sup>3</sup>, Ahmed Salem<sup>4</sup>

College of Information Technology, Misr University for Science and Technology, Giza, Egypt<sup>1,3</sup>

Faculty of Engineering, Al-Azhar University, Cairo, Egypt<sup>2</sup>

College of Computing & Information Technology, Arab Academy for Science, Technology & Maritime Transport, Cairo, Egypt<sup>4</sup>

**Abstract**—Multi-Criteria Decision-Making (MCDM) techniques are often used to aid decision-makers in selecting the best alternative among several options. However, these systems have issues, including the Rank Reversal Problem (RRP) and decision-making ambiguity. This study aimed to propose a selection model for a Cloud Service Provider (CSP) that addresses these issues. This research used the Technique for Order Preference by Similarity to Ideal Solution (TOPSIS) to rank the alternatives. The entropy technique is utilized to determine the weight of the criteria, and Single Valued Neutrosophic (SVN) is employed to address uncertainty. To select the best cloud provider based on Quality of Service (QoS) criteria, we used a dataset from Cloud Harmony for this study. The results indicated that the suggested model could effectively resolve the RRP under conditions of uncertainty. This research is novel and is the first to address both the problem of uncertainty in decision-making and RRP in MCDM.

**Keywords**—MCDM; TOPSIS; neutrosophic set; single valued neutrosophic; cloud services provider; quality of service

## I. INTRODUCTION

Cloud computing, an emerging paradigm, offers users pay-per-use or on-demand services. It provides users with three primary categories of service models: infrastructure as a service (IaaS), platform as a service (PaaS), and software as a service (SaaS). IaaS offers computational assistance to cloud clients. PaaS provides a framework for application development. SaaS gives users access to pre-made apps. Due to the vast number of software products and flexibility in utilizing cloud services, many large firms, such as Microsoft and Google, are investing significant amounts of money in offering various cloud services. However, finding and identifying a CSP has become a challenging task for cloud users due to the growing number of cloud providers.

Cloud benchmarking service providers, such as Cloud Harmony and Cloud Spectator [1, 2], analyze the performance of multiple CSPs and publish their findings online, serving as the foundation of the simple method cloud customers use to select the optimal CSP. However, the execution environment used by cloud customers may differ from the performance assessed by a third party in a given context. As a result, professionals or cloud users must evaluate multiple CSPs based on their experience to choose the optimal CSP.

The above issue has motivated researchers to design a mechanism for selecting the optimal CSP, which requires a set of QoS criteria to assess cloud services and a methodology for rating them according to these criteria [3].

MCDM is a structured and formal decision-making approach used to deal with complex problems and conflicting criteria.

There are several MCDM approaches used in related works, such as TOPSIS, Decision-Making Trial and Evaluation Laboratory (DEMATEL), Simple Additive Weightage (SAW), VIseKriterijumska Optimizacija I Kompromisno Resenje (VIKOR), Analytic Hierarchy Process (AHP), Elimination Et Choice Translating Reality (ELECTRE), Preference Ranking Organization Method for Enrichment Evaluations (PROMETHEE), Complex Proportion Assessment Method (COPRAS), Analytic Network Process (ANP), Multi-objective Optimization on the Basis of Ratio Analysis Method (MOORA), Stepwise Weight Assessment Ration Analysis (SWARA), and others [4], [5], [6], [7] and [8].

Generally, TOPSIS is the most popular technique for handling MCDM problems [9]. It depends on synthesizing the criteria and dividing the alternatives into two subsets: positive and negative solutions. The optimal solution has the shortest distance from the positive set of solutions and the longest distance from the negative set of solutions [10].

Due to its advantages over other fuzzy extensions, SVN Set has been taken into consideration for handling vagueness. The membership function, used in fuzzy set theory developed by Zadeh et al. [11], translates linguistic terms into membership values. However, the value of membership for a term may vary among experts. For example, one expert may give a value of two to express the linguistic term "low," while another may give a value of three.

To address this issue, a Neutrosophic Set (NS) is employed. With the condition that the three membership values must be less than or equal to three, NS allows decision-makers to judge in three degrees: truth, indeterminacy, and falsity. As a generalization of all fuzzy set versions, NS has been combined with several MCDM techniques and aids decision-makers in resolving ambiguity in their judgment [12].

The main contributions of this research can be summarized as follows:

- A comprehensive analysis of the robustness of MCDM models against the RRP.
- An evaluation of the suggested model's resistance to RRP.
- The integration of SVN theory with a modified Entropy-based TOPSIS method.
- A comparative analysis between the proposed model and other MCDM models.

The rest of the paper is organized as follows: Section II discusses related work. Section III presents the methodology used in the proposed model. In Section IV, we present the results and validation of our model. Section V summarizes the conclusions of our research, and Section VI outlines future work.

## II. RELATED WORK

The increasing number of CSPs has attracted the interest of researchers in evaluating their performance in different applications. The primary objective of this research is to assess CSP performance and develop techniques for finding the most effective and optimal CSP. MCDM techniques have been extensively utilized in previous publications to handle decision-making problems in various industries, such as supplier and employee selection. Since the current proposed methodology combines TOPSIS with NS to identify the optimal CSP, we first explored MCDM-based techniques. Then we reviewed numerous publications that used NS in conjunction with MCDM to tackle different decision-making problems. After that, we highlighted the drawbacks of the MCDM-based TOPSIS technique.

Zulqarnain et al. [13] applied neutrosophic TOPSIS to select the most suitable supplier and found that neutrosophic can handle uncertainty in decision-making. However, they did not consider the RRP in TOPSIS.

Garcia et al. [14] discovered that the TOPSIS technique suffers from RRP due to changing the normalized value of the judgment matrix when an alternative is added. They proposed two hypothetical values representing the minimum and maximum values for each criterion, and the modified technique can handle some cases of RRP.

Abdel-Basset et al. [15] developed a hybrid technique combining neutrosophic set theory and AHP to evaluate cloud services. They implemented a function to convert linguistic terms into crisp values. The hybrid technique is effective when classical AHP fails due to an inconsistent pairwise decision matrix; however, it does not address RRP.

Kumar et al. [16] developed a hybrid technique by combining AHP and TOPSIS. AHP is used to obtain each criterion's weight, and TOPSIS is used to rate CSPs based on cloud benchmarking reports. A significant limitation of this research is that it cannot handle the uncertainty problem, or RRP, in MCDM.

Jatoth et al. [17] developed an integrated model that consists of AHP and grey TOPSIS. The grey set is used to handle uncertainty in decision-making. The proposed model considers both functional and non-functional requirements of cloud services but does not consider the RRP.

Aires et al. [18] proposed R-TOPSIS, a modified version of TOPSIS. This model requires a judgment matrix, criteria weights, and domains for each criterion. It uses the domain of each criterion with a max or max-min normalization approach to normalize the judgment matrix. The Positive Ideal Solution (PIS) and Negative Ideal Solution (NIS) are computed using a novel method. The results showed that the model fails to handle RRP when removing a non-distinct criterion.

Goswami et al. [19] proposed a technique for choosing the optimal steel grades and their corresponding heat treatment procedures using a hybrid technique based on entropy and TOPSIS. The limitation of this model is that it cannot handle uncertainty or RRP.

Tiwari et al. [12] developed a framework based on neutrosophic TOPSIS to handle uncertainty in decision-making. The framework is validated against only two types of RRP: the insertion and deletion of alternatives from the decision matrix.

Hezam et al. [20] developed an MCDM model based on neutrosophic AHP and TOPSIS to identify the priority groups for the COVID-19 vaccine. The model was able to handle uncertainty, but it has not been validated against the RRP.

Trabay et al. [21] built a mathematical model based on MCDM to rate the trustworthiness of cloud services based on various opinions. The results showed that fuzzy TOPSIS provides more accurate results than TOPSIS, fuzzy AHP, and AHP.

Saha et al. [22] proposed a hybrid MCDM model consisting of ANP and VIKOR, where ANP is used to obtain the local rank of CSP, and VIKOR is used to obtain the global rank. The major disadvantage of this model is that it cannot handle uncertainty or RRP.

Dani et al. [23] developed a technique to assess the efficiency of educational boards. They used a linear weighted model and TOPSIS. The results showed that the ranks obtained by both models were very similar.

Dhand et al. [24] developed a network selection model consisting of fuzzy AHP and ELECTRE, where fuzzy AHP is utilized to obtain the weight of each criterion, and ELECTRE is utilized to rate networks. Results showed that the model could effectively select the optimal network, but it has not been validated against the RRP.

According to previous research, we can consider CSPs ranking as a decision problem. The majority of researchers employed MCDM to select the optimal CSP. Some techniques are extended to fuzzy or NS theories to handle uncertainty. The previously discussed related works addressed either the RRP or uncertainty, but none tried to address both rank reversal and uncertainty simultaneously.

NS has become essential in solving decision problems because it can more efficiently handle uncertainty problems in decision-making. Therefore, we used a neutrosophic set with an integrated Entropy-TOPSIS technique to choose the optimal CSP. The novel model is effective and robustly selects CSPs in the neutrosophic state. Our research is the first to apply the integrated Entropy-TOPSIS technique to CSP ranking.

### III. METHODOLOGY

#### A. Basic Concepts

This section introduces Entropy, TOPSIS and some basic definitions of NS and SVN.

1) *Neutrosophic set theory*: This theory considers every idea  $\langle X \rangle$  along with its negation  $\langle \text{Anti-}X \rangle$  and a group of "neutralities,"  $\langle \text{Neut-}X \rangle$ , which lies between the two boundaries and supports neither  $\langle X \rangle$  nor  $\langle \text{Anti-}X \rangle$  [11].

a) *Single valued neutrosophic set*: Let  $X$  be a space of objects,  $x \in X$ . A neutrosophic set  $N$  on  $X$  is defined by a truth membership  $T_N$ , an indeterminacy membership  $I_N$ , and a falsity membership  $F_N$ .  $T_N(x)$ ,  $I_N(x)$  and  $F_N(x)$  are subsets of  $[0, 1]^+$ , and the sum of their values is between  $0$  and  $3^+$  [12].

b) *Score function*: Junaid et al. [25] proposed the following score functions  $S(x_{i,j})$  to transform the neutrosophic numbers into crisp numeric value.

$$S(x_{i,j}) = \frac{Lx_{i,j} + Mx_{i,j} + Ux_{i,j}}{3} + (T_{x_{i,j}} - I_{x_{i,j}} - F_{x_{i,j}}) \quad (1)$$

$$S(x_{i,j}) = \frac{1}{s(x_{i,j})} = \frac{1}{\frac{Lx_{i,j} + Mx_{i,j} + Ux_{i,j}}{3} + (T_{x_{i,j}} - I_{x_{i,j}} - F_{x_{i,j}})} \quad (2)$$

Where  $L$ ,  $M$ , and  $U$  are the lower, medium, and upper values of the neutrosophic numbers, and  $T$ ,  $I$ , and  $F$  are the degrees of truthiness, indeterminacy, and falsity. If there is more than one decision expert, then the average of all experts' scores should be calculated to obtain the aggregated matrix [25].

2) *Entropy*: Entropy is an objective weighting method developed by Shannon [26]. It is used to calculate the weight of criteria for a multi-objective decision problem without considering the decision-makers' opinions. Weights are identified using the entropy method, which automatically computes the weight of criteria based on the judgment matrix, i.e., the significance of the parameter in relation to the other parameters. The steps of entropy are listed in [19].

3) *TOPSIS*: TOPSIS is the most widely used MCDM method, which ranks alternative solutions based on increasing the distance from the negative ideal point and reducing the distance from the positive ideal point. The steps involved in the TOPSIS method are listed in [19].

#### B. Proposed Model

Aires et al. [18] determined that an effective solution for the RRP in TOPSIS technique should take the following factors into account at the same time:

- Selecting a normalization method that reduces the consequences of alternative dependence.
- Using fixed NIS and PIS even if the set of alternatives is modified.

In addition to that, a lot of related works used the neutrosophic set to eliminate uncertainty in decision-making [25]. Therefore, we proposed a model based on SVN numbers to handle uncertainty problems, and we modified the normalization procedure in the Original TOPSIS technique. Moreover, a normal Gaussian distribution for normalization [27] and fixed PIS and NIS were used to calculate the rank of alternatives.

The steps of the proposed model are given in Algorithm 1, and a schematic diagram of the proposed model is presented in Fig. 1.

---

#### Algorithm 1: Proposed Model

---

##### A. Phase I: Modified Entropy

**Input:** The decision matrix  $D$  ( $m \times n$ ) which contains the performance values and is represented in linguistic terms. 'm' denotes the number of alternatives, and 'n' denotes the number of criteria.

**Output:** The weight of each criterion.

*Step 1:* Create a Decision Matrix  $D$ .

$$D = \begin{bmatrix} x_{1,1} & \dots & x_{1,n} \\ \vdots & \ddots & \vdots \\ x_{m,1} & \dots & x_{m,n} \end{bmatrix} \quad (3)$$

*Step 2:* Map each linguistic term to its equivalent SVN value using Table I.

TABLE I. MAPPING BETWEEN LINGUISTIC TERM AND SVNS

Linguistic terms	SVNs
Extremely Low (EL)	$\langle (1, 2, 3); (0.3, 0.75, 0.7) \rangle$
Very Low (VL)	$\langle (2, 3, 4); (0.4, 0.6, 0.65) \rangle$
Low (L)	$\langle (3, 4, 5); (0.6, 0.35, 0.4) \rangle$
Medium Low (ML)	$\langle (4, 5, 6); (0.7, 0.3, 0.35) \rangle$
Medium/Fair (M/F)	$\langle (1, 1, 1); (0.5, 0.5, 0.5) \rangle$
Medium High (MH)	$\langle (5, 6, 7); (0.8, 0.25, 0.3) \rangle$
High (H)	$\langle (6, 7, 8); (0.85, 0.2, 0.25) \rangle$
Very High (VH)	$\langle (7, 8, 9); (0.9, 0.15, 0.2) \rangle$

Step 3: Convert the SVN into crisp numbers using score function given in Equation 1.

Step 4: Compute normalized decision matrix ‘ $r_{i,j}$ ’.

$$r_{i,j} = \frac{x_{i,j}}{\sum_{i=1}^m x_{i,j}} \quad (4)$$

Step 5: Calculate the entropy value ‘ $e_j$ ’ for each criterion.

$$e_j = -h \sum_{i=1}^m r_{i,j} \ln(r_{i,j}) \quad (5)$$

where  $h=1/\ln(m)$  and  $m$  is the number of alternatives.

Step 6: Compute the degree of divergence ‘ $G_j$ ’ for each criterion.

$$G_j = |1 - e_j| \quad (6)$$

Step 7: Compute the weight of each criterion ‘ $w_j$ ’.

$$w_j = \frac{G_j}{\sum_{j=1}^n G_j} \quad (7)$$

### B. Phase II: Modified TOPSIS

**Input:** The same decision matrix in (phase I) and the criterion weightages ‘ $w_j$ ’ from (phase. I).

**Output:** The rank of each alternative.

Step 1: Calculate the normalized decision matrix ‘ $M$ ’ using the normal Gaussian distribution function  $F(x_{i,j})$ .

$$m_{i,j} = F(x_{i,j}) = \int_{-\infty}^{x_{i,j}} e^{-(x-\mu)^2/2\sigma^2} dx \quad (8)$$

Step 2: Calculate the weighted normalized decision matrix.

$$W_{i,j} = w_j * m_{i,j} \quad (9)$$

Step 3: Calculate the PIS and NIS using the following equations.

$$v_j^+ = \{v_1^+ \dots v_m^+\} = \begin{cases} v_j^+ = w_j & \text{if } j \in \text{Benefit Criteria} \\ v_j^+ = 0 & \text{if } j \in \text{Cost Criteria} \end{cases} \quad (10)$$

$$v_j^- = \{v_1^- \dots v_m^-\} = \begin{cases} v_j^- = 0 & \text{if } j \in \text{Benefit Criteria} \\ v_j^- = w_j & \text{if } j \in \text{Cost Criteria} \end{cases} \quad (11)$$

Step 4: Compute the Euclidean distance  $S_i^+$  and  $S_i^-$  of each alternative from the PIS and NIS.

$$S_i^+ = \left[ \sum_{j=1}^m (v_{i,j} - v_j^+)^2 \right]^{1/2} \quad (12)$$

$$S_i^- = \left[ \sum_{j=1}^m (v_{i,j} - v_j^-)^2 \right]^{1/2} \quad (13)$$

Step 5: Calculate the closeness index ( $P_i$ ) for each alternative.

$$P_i = \frac{S_i^-}{S_i^+ + S_i^-} \quad (14)$$

Step 6: Rank each alternative based on its relative closeness index ( $P_i$ ) in descending order.

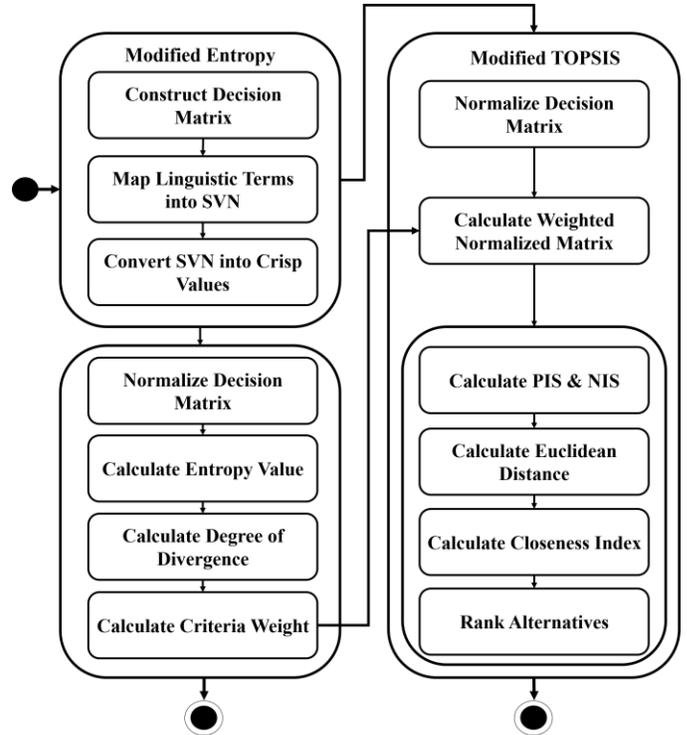


Fig. 1. Schematic diagram of the proposed model

## IV. CASE STUDY

The proposed CSP selection methodology assists cloud users in choosing the appropriate cloud service for their needs. A case study was carried out to validate its reliability, where we applied the proposed model using a real dataset obtained from Tiwari et al. [12]. This dataset was obtained from reports issued by Cloud Harmony [1], a cloud benchmarking service provider. In addition to applying our proposed model in the case study, we conducted sensitivity and performance studies on the proposed model. Furthermore, the model was compared and validated with other studies.

### A. Data Set

The dataset contains ten QoS parameters from six real-world CSPs. The QoS parameters used are Cost (C), Latency of Network (NL), Sequential Disk RW Performance Consistency (SDRWPC), Random Disk RW Performance Consistency (RDRWPC), CPU Integer Performance (CPUIP), CPU Floating Point Performance (CPUFPP), Memory Performance on Scale (MPS), Memory Performance on Triad (MPT), Sequential RW Disk Performance (SRWDP) & Random RW Disk Performance (RRWDP). The first four criteria are costly, while the others are benefit criteria. Table II shows the dataset, where all values are represented in linguistic terms [12].

TABLE II. DATA SET [12]

CSP	C	NL	SDRWPC	RDRWPC	CPUIP	CPUFPP	MPS	MPT	SRWDP	RRWDP
Soft Layer	L	VL	L	F	L	L	H	H	F	F
Rack Space	F	F	F	F	L	L	H	H	H	L
Ms. Azure	L	F	L	L	VL	VL	F	F	L	L
Google	L	H	L	L	L	L	H	H	VL	VL
Digital Ocean	L	F	VL	VL	L	L	H	H	L	L
Amazon EC2	L	VL	L	L	L	L	H	H	L	L

### B. Cloud Service Ranking

The proposed model was used to calculate the rank of CSPs. Table II presents the judgment matrix for the dataset used in this study, which was formed based on 10 QoS measures and 6 CSPs. After applying the proposed model shown in (Algorithm 1), the rank of each CSP was calculated using the closeness index. Rack space was ranked as the optimal CSP, while Amazon EC2 was the worst CSP.

### C. Sensitivity Analysis

The analysis was conducted to evaluate the proposed model's reliability and consistency in different RRP scenarios. This analysis had two objectives: the first was to determine the RRP when changing the number of alternatives, and the second was to test the proposed model's reliability. This was achieved by performing a complete test to observe a variation in each case. The ranking model is considered to be reliable if it consistently ranked CSPs.

The Five types of RRP are discussed as follows:

1) *The first type: Deletion of an alternative from the dataset:* This rank reversal analysis was carried out by deleting only one CSP. Six experiments were performed on the Cloud Harmony dataset (Table. 2). In each experiment, a single CSP was deleted. No changes were observed in the closeness index and rank of alternatives, demonstrating that the proposed model is resistant to the RRP found in the first type.

2) *The second type: Addition of an alternative to the dataset:* This rank reversal analysis was carried out by adding a CSP alternative. Four experiments were performed, and the closeness index and rank were calculated each time. The rank of alternatives was not affected by adding any alternatives, demonstrating that the proposed model is resistant to the RRP found in the second type.

3) *The third type: Addition of an irrelevant alternative:* The third type of RRP is carried out by adding an irrelevant alternative to the dataset to assess the reliability of the proposed model. We added an irrelevant  $(CSP)_x$  with the same criteria as Rack Space alternative which exists in the data set (Table II). The Closeness index and rank were computed, and

we observed that rank was the same before adding an irrelevant alternative  $(CSP)_x$ .

4) *The fourth type: Testing the property of transitivity by dividing the existing matrix into two sub-decision matrices:* In the Fourth type, the decision matrix was divided into two subsets, then the rank was calculated in each sub set and compared to the rank of the original decision matrix. After performing this test, we found out that the rank obtained from the two subsets was the same as the rank obtained from the original matrix.

5) *The fifth type: The deletion of a non-distinct criterion:* The Fifth type was carried out by deleting a non-distinct criterion from the existing dataset. A non-distinct criterion is a criterion with the lowest standard deviation. In our experiment the third criterion named SDRWPC was removed since it had the lowest standard deviation, then we observed that rank obtained after removing the non-distinct criterion was the same as the rank obtained before removing this criterion, which demonstrates that the proposed model is resistant to RRP found in type five.

The above rank reversal sensitivity analysis for all test cases showed that the proposed model is resistant to the RRP.

### D. Results Validation

A validation was conducted to verify the accuracy of the rank calculated by the proposed model as follows:

1) *Firstly,* comparative analysis was performed to validate the proposed model. Fig. 2 demonstrates the ranking of each CSP obtained using the model and related work. Rack space was ranked first in all techniques, Google second, Digital Ocean third, Ms. Azure fourth, and soft layer and Amazon EC2 sixth and fifth, respectively, in all techniques except the technique proposed by Kumar et al. [16]. In addition, the model ranked the alternatives almost identically to those developed by Goswami et al. [19] and Aires et al. [18]. In contrast, it slightly differed from the model developed by Kumar et al. [16]. Therefore, it could be concluded that the proposed model accurately ranks CSPs.

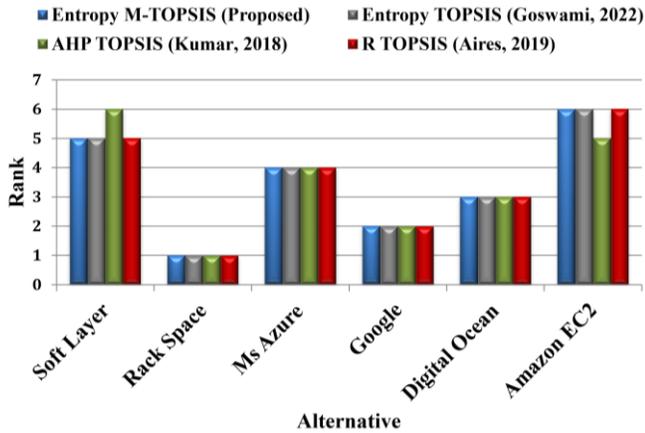


Fig. 2. The rank of CSP for the proposed model & related work

2) *Secondly*, the suggested model and other related work were compared [18], [19], and [16]. Table III shows the resistance to the RRP for the proposed model and related work. The proposed model could handle all five types of RRP, while the original TOPSIS, Entropy-TOPSIS, and AHP-TOPSIS techniques could not handle all kinds of RRP. R-TOPSIS, developed by Aires et al. [18], could handle all types of RRP except type five.

TABLE III. THE RRP ANALYSIS FOR THE PROPOSED MODEL AND RELATED WORK

Method	Type 1	Type 2	Type 3	Type 4	Type 5
Proposed Model	√	√	√	√	√
Entropy-TOPSIS	X	X	X	X	X
AHP-TOPSIS	X	X	X	X	X
R-TOPSIS	√	√	√	√	X
TOPSIS	X	X	X	X	X

3) *Thirdly*, statistics of dispersion and similarity were used to compare the techniques in related work with the proposed model. The following statistical methods were applied in validation phase [18], [28] and [29]:

- *Similarity*: Mean Absolute Error of Rank (MAER) and Spearman’s Rank Correlation (SRC).
- *Dispersion*: The Standard Deviation of the closeness coefficient for each rank (SD), the difference between the closeness coefficient of the best alternative and the worst (BWD), and the difference between closeness coefficient of the 1<sup>st</sup> and 2<sup>nd</sup> alternative (FSD).

The simulation was implemented in MATLAB. Table IV compares the proposed model and its other variants, considering similarity (SRC and MAER) statistics and dispersion (SD, BWD and FSD) statistics. An average of four simulation cycles was used for 4,000 simulated cases.

Comparing the dispersion measures for each technique, the proposed model generally had smaller values for (SD, BWD and FSD) than the other models. Furthermore, based on the statistical methods used to compute the similarity degree between the ranks obtained by the four techniques, it was observed that there is a very high degree of similarity between the rankings, indicating that the suggested model corresponds to the other methods. Fig. 3 and Fig. 5 show that the MAER value between the proposed model (M1) and method (M2) is the lowest. In contrast, Fig. 4 and Fig. 6 show that the SRC value between (M1) and (M2) is the highest. Moreover, method (M4) deviated from the proposed model more than the other methods. Therefore, it can be concluded that the proposed model is more similar to the method (M2) according to all similarity measures.

TABLE IV. DISPERSION AND SIMILARITY STATISTICS

Method	SD	BWD	FSD	SRC	MAER
Entropy M-TOPSIS (M1)	0.0582	0.1106	0.0569	0.8120	0.0823
Entropy TOPSIS (M2)	0.1959	0.3740	0.1946		
Entropy M-TOPSIS (M1)	0.0582	0.1106	0.0569	0.2295	0.3420
AHP-TOPSIS (M3)	0.2109	0.3989	0.1946		
Entropy M-TOPSIS (M1)	0.0582	0.1106	0.0569	0.1080	0.3963
R TOPSIS (M4)	0.1504	0.2858	0.1674		

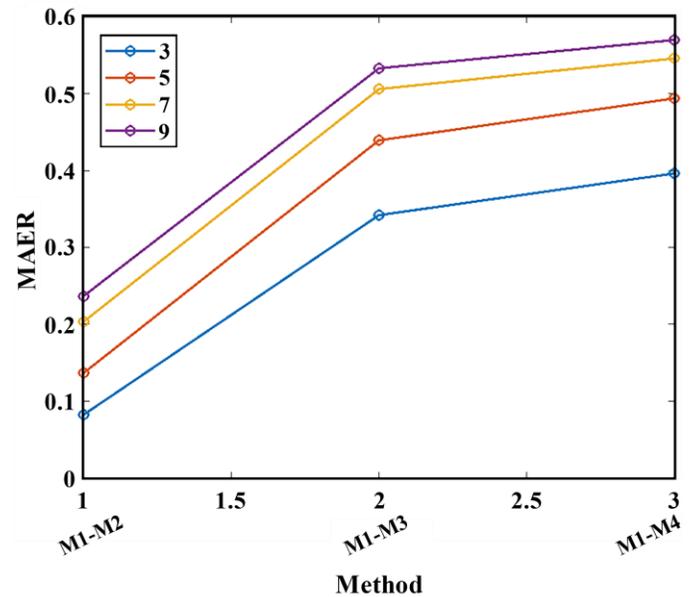


Fig. 3. MAER by number of alternatives

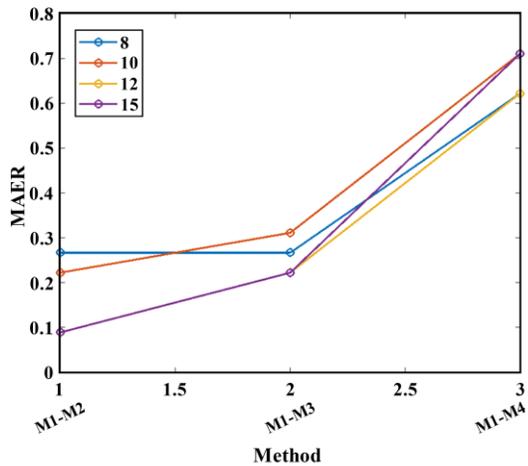


Fig. 4. MAER by number of criteria

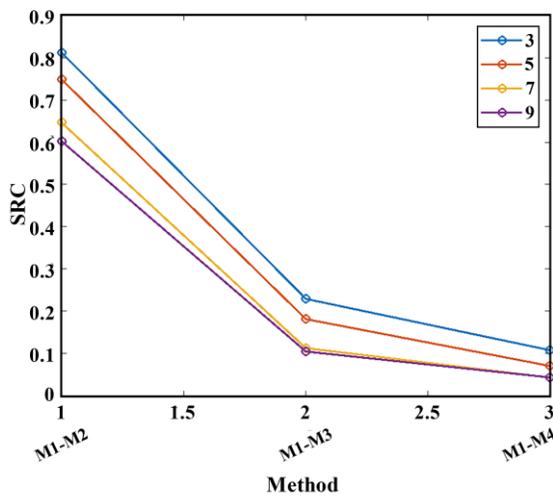


Fig. 5. SRC by number of alternatives

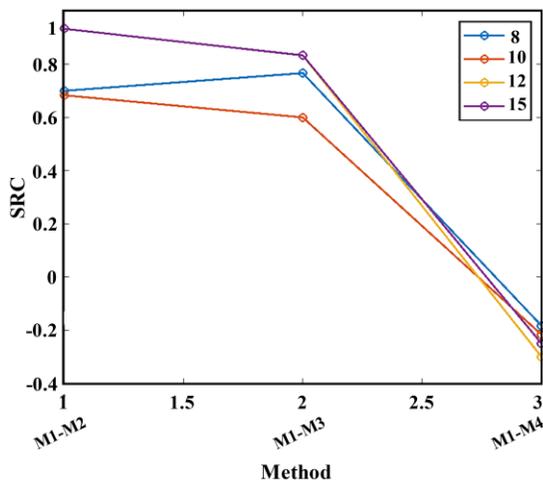


Fig. 6. SRC by number of criteria

4) *Fourthly*, the execution time of the proposed model and related work was measured with increasing the number of alternatives. The analysis was carried out using a Core i5 (8<sup>th</sup> Gen.) PC, with WIN 10 (64-bit) OS, and 8 G.B RAM. Linguistic terms for about 1,500 alternatives and ten criteria were randomly generated. Fig. 7 shows the execution time of the proposed model and other related work. It can be noted The proposed model took less time to execute than Entropy-TOPSIS and AHP-TOPSIS and only slightly more time than R-TOPSIS, by just a few milliseconds.

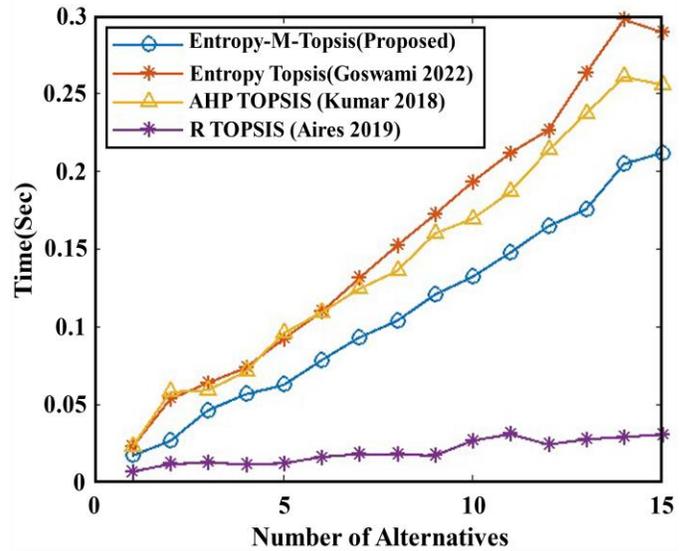


Fig. 7. Analysis of the execution time

## V. CONCLUSION

The analysis of RRP in TOPSIS presented in this study is comprehensive. We have identified that the normalization process and the selection of PIS and NIS are the leading causes of RRP in TOPSIS. To address these issues, we utilized the normal Gaussian distribution function for normalization and introduced a new approach for calculating PIS and NIS. Moreover, we proposed a novel extension to handle insufficient information, including degrees of truth, indeterminacy, and falsity, by integrating SVN with the suggested model. The proposed model was validated through sensitivity analysis, comparative analysis, and statistical measures of similarity and dispersion. The results indicated that the proposed model could improve the decision-making process under uncertainty with high accuracy and robustness against RRP, making it applicable to any multi-criteria decision problem. One limitation of this research is that it did not consider subjective weighting-based approaches that determine criteria weights based on the judgments of decision-makers. Table V summarizes the overall results of this research work.

TABLE V. SUMMARY OF THE OVERALL RESULTS

Model	Evaluation Criteria	
	Handling Uncertainty	Handling All Types of RRP
Proposed model	√	√
Entropy-TOPSIS	x	x
R-TOPSIS	x	x
AHP-TOPSIS	x	x

## VI. FUTURE WORK

New extensions of neutrosophic sets can be utilized to solve the uncertainty problem and provide more accurate results to support the decision-making process. In addition, the output from various MCDM techniques, such as COPRAS, PROMETHEE, and others, can be compared to the results of the current research. Finally, other subjective and objective weighting-based approaches can be utilized, and the difference in rank can be assessed.

## ACKNOWLEDGMENT

We would like to thank Prof. F. Smarandache and Dr. Mohamed Abdel-Basset, for their support.

## REFERENCES

- [1] "Cloud Harmony," [Online]. Available: <https://cloudharmony.com/> [Accessed 01.06.2021].
- [2] "Cloud Spectator," [Online]. Available: <https://cloudspectator.com/> [Accessed 01.06.2021].
- [3] D. Ardagna, G. Casale2, M. Ciavotta, J. F. Pérez and W. Wang, "Quality-of-service in cloud computing: modeling techniques and their applications," *Internet Services & Apps*, Vol.5, No.11, pp.1-17, 2014.
- [4] S. Mufazzal and S.M. Muzakkir, "A new multi-criterion decision-making MCDM method based on proximity indexed value for minimizing rank reversals," *Computers & Industrial Engineering*, Vol.118, pp.427-438, 2018.
- [5] J. J. Thakkar, "Studies in Systems, Decision and Control," in Multi-Criteria Decision-Making, *Springer Singapore*, Vol.336, pp.1-390, 2021.
- [6] A. Alghawli, A. Al-khulaidi, N. Al-khulaidi, A. Nasser, and F. Abass, "Application of the Fuzzy Delphi Method to Identify and Prioritize the Social-Health Family Disintegration Indicators in Yemen," *International Journal of Advanced Computer Science and Applications (IJACSA)*, Vol.13, No.5, pp.680-691, 2022.
- [7] B. Sahoo, R. Behera, P. Pattnaik, "A Comparative Analysis of Multi-Criteria Decision-Making Techniques for Ranking of Attributes for e-Governance in India," *International Journal of Advanced Computer Science and Applications (IJACSA)*, Vol.13, No.3, pp.65-70, 2022.
- [8] A. Arafat, L. Beh, "Factors Influencing Practice of Human Resource Information System in Organizations: A Hybrid Approach of AHP and DEMATEL," *International Journal of Advanced Computer Science and Applications (IJACSA)*, Vol.11, No.6, pp.700-706, 2020.
- [9] E. K. Zavadskas, A. Mardani, Z. Turskis, A. Jusoh and K. M. Nor, "Development of TOPSIS Method to Solve Complicated Decision-Making Problems: An Overview on Developments from 2000 to 2015," *International Journal of Information Technology & Decision-Making*, Vol. 15, No.3, pp.645-682, 2016.
- [10] N. A. Nabeeh, F. Smarandache, M. Abdel-Basset, H. A. El-Ghareeb and A. Aboelfetouh, "An integrated neutrosophic-TOPSIS approach and its application to personnel selection: A new trend in brain processing and analysis," *IEEE Access*, Vol.7, pp.29734--29744, 2019.
- [11] S. Broumi, A. Bakali, M. Talea, F. Smarandache, V. Ulucay, M. Sahin, A. Dey, M. Dhar, R. Tan, A. Bahnasse and S. Pramanik, "Neutrosophic Sets: An Overview," in *New Trends in Neutrosophic Theory and Applications*, Vol.2, pp.388-418, 2018.
- [12] R. K. Tiwari and R. Kumar, "A framework for prioritizing cloud services in neutrosophic environment," *Journal of King Saud University-Computer and Information Sciences*, Vol.34, No.6, pp.3151-3166, 2022.
- [13] R. M. Zulqarnain, X. L. Xin, M. Saqlain, F. Smarandache and M. Irfan, "An integrated model of neutrosophic TOPSIS with application in multi-criteria decision-making problem," *Neutrosophic Sets and Systems*, Vol.40, pp.253-269, 2021.
- [14] M. S. García-Cascales and M. T. Lamata, "On rank reversal and TOPSIS method," *Mathematical and Computer Modelling*, Vol.56, No.5, pp.123-132, 2012.
- [15] M. Abdel-Basset, M. Mohamed and V. Chang "NMCD: A framework for evaluating cloud computing services," *Future Generation Computer Systems*, Vol.86, pp.12-29, 2018.
- [16] R. R. Kumar, S. Mishra and C. Kumar, "A novel framework for cloud service evaluation and selection using hybrid MCDM methods," *Arabian Journal for Science & Engineering*, Vol.43, pp.7015-7030, 2018.
- [17] C. Jatoth, G. R. Gangadharan, U. Fiore and R. Buyya, "SEL-CLOUD: a hybrid multi-criteria decision-making model for selection of cloud services," *Soft Computing*, Vol.23, pp.4701-4715, 2019.
- [18] R. F. d. F. Aires and L. Ferreira, "A new approach to avoid rank reversal cases in the TOPSIS method," *Computers & Industrial Engineering*, Vol.132, pp.84-97, 2019.
- [19] S. S. Goswami, S. Jena and D. K. Behera, "Selecting the best AISI steel grades and their proper heat treatment process by integrated entropy-TOPSIS decision-making techniques," *Materials Today: Proceedings*, Vol.60, pp.1130-1139, 2022.
- [20] I. M. Hezam, M. K. Nayeem, A. Foul and A. F. Alrasheedi, "COVID-19 Vaccine: A neutrosophic MCDM approach for determining the priority groups," *Results in Physics*, Vol.20, pp.103654, 2021.
- [21] D. Trabay, A. Asem, I. El-Henawy and W. Gharibi, "A hybrid technique for evaluating the trust of cloud services," *Int. j. of Information Technology*, Vol.13, No.2, pp.687-695, 2021.
- [22] M. Saha, S. K. Panda and S. Panigrahi, "A hybrid multi-criteria decision-making algorithm for cloud service selection," *Int. j. of Information Technology*, Vol.13, No.4, pp.1417-1422, 2021.
- [23] D. Dani and G. Agrawal, "Evaluating the Quality of Indian School Education boards' websites using multi-criteria decision-making models," *Int. j. of Information Technology*, Vol.13, No.6, pp.2551-2559, 2021.
- [24] P. Dhand, S. Mittal and G. Sharma, "An intelligent handoff optimization algorithm for network selection in heterogeneous networks," *Int. j. of Information Technology*, Vol.13, No.5, pp.2025-2036, 2021.
- [25] M. Junaid, Y. Xue, M. W. Syed, J. Z. Li and M. Ziaullah, "A Neutrosophic AHP and TOPSIS Framework for Supply Chain Risk Assessment in Automotive Industry of Pakistan," *Sustainability*, Vol.12, No.1, pp.154, 2020.
- [26] C. Bhowmik, M. A. Kaviani, A. Ray and L. Ocampo, "An integrated entropy-TOPSIS methodology for evaluating green energy sources," *International Journal of Business Analytics*, Vol.7, No.3, pp.44-70, 2020.
- [27] A. Z. Sarraf, A. Mohaghar and H. Bazargani, "Developing TOPSIS method using statistical normalization for selecting Knowledge management strategies," *Journal of Industrial Engineering and Management*, Vol.6, No.4, pp.860-875, 2013.
- [28] I. Chamodrakas, I. Leftheriotis and D. Martakos, "In-depth analysis and simulation study of an innovative fuzzy approach for ranking alternatives in multiple attribute decision-making problems based on TOPSIS," *Applied Soft Computing*, Vol.11, No.1, pp.900-907, 2011.
- [29] L. Ferreira, D. Borenstein and E. Santi, "Hybrid fuzzy MADM ranking procedure for better alternative discrimination," *Engineering Applications of Artificial Intelligence*, Vol.50, pp.71-82, 2016.

# Dynamic Software Architecture Design for Virtual Rehabilitation System for Manual Motor Dexterity

Edwin Enrique Saavedra Parisaca<sup>1</sup>, Solansh Jaqueline Montoya Muñoz<sup>2</sup>, Elizabeth Vidal Duarte<sup>3</sup>, Eveling Gloria Castro Gutierrez<sup>4</sup>, Angel Yvan Choquehuanca Peraltilla<sup>5</sup>

Universidad Nacional de San Agustín de Arequipa  
Arequipa, Perú

Sergio Albiol Pérez<sup>6</sup>  
Universidad Nacional de Zaragoza  
Teruel, España

**Abstract**—The architectural design is fundamental in the construction process of a virtual rehabilitation system since it allows to understand the components and their interaction, and it is a guide to develop the software. This article proposes a dynamic architecture design that could be used independently of software and hardware in a virtual rehabilitation system for motor dexterity. This proposal contributes to the software engineering area since it provides a starting point for the development of virtual rehabilitation systems. The system implementation was done with two tracking devices (hardware) and two rehabilitation games (software). It was validated with the User Experience Questionnaire (UEQ). Results show that the use of a dynamic architecture allowed to use different devices efficiently and quickly, regardless of the game, preventing the user from feeling a change or difficulty in carrying out the tasks.

**Keywords**—Software architecture; dynamic architecture; virtual rehabilitation systems; motor dexterity

## I. INTRODUCTION

Software architecture as a subfield in engineering defines the structure, operation, properties and relationship between software elements [1, 2].

Today the changing nature of technology has driven the customization and application of modern design techniques to meet the software requirements of emerging systems [2], one of them is Virtual Rehabilitation Systems (RhV) [3]. These types of systems have simplified the need for physical environments and have shown good results in the area of rehabilitation through repetition, sensory feedback, multisensory stimulation, and motivation [4].

Based on the literature review [5-11] it was seen that existing designs based on proposed software architectures for RhV systems are poorly described or difficult to maintain and evolve over time.

Even seeing that the recurring variation between the different proposals [5-11] is that its components, despite maintaining the same functional purpose, are continually added, eliminated, and replaced due to the characteristics of the patients, budgetary limitations, or the variety of options available that allow the development of these systems, such as Tech-MCS for remote control of therapy [12], CyberTouch or SaebOGlove applied to the rehabilitation of upper extremities [13, 14], Leap motion, among others.

Giving a system the ability to adapt to these aspects that are already structural is a case of application of a Dynamic Architecture. A type of architecture where there is variation between the number of nodes and links that compose them topologically [15].

It is essential to encourage this type of architecture so that the systems evolve and adapt since there are various diseases that cause problems in motor dexterity such as epileptic encephalopathy, which presents abnormalities associated with progressive brain dysfunction and causes motor impairment due to epileptic activity generated by seizures [16]. Another example is Acquired Brain Injury (ACD) which produces physical, neurocognitive, and/or psychological deficiencies [17]. Its fundamental consequence is the loss of previously developed brain functions that involve the motor and sensory systems [18].

Other examples are autism, cerebral palsy, and stroke, among others. All these conditions cause motor difficulties in the daily life of those who suffer from them, becoming severely disabling on many occasions and requiring rehabilitation tasks. According to the World Health Organization (WHO) [19] 16% of the world's population – experiences significant disability today.

The objective of the work was the design of a dynamic software architecture for virtual rehabilitation systems to improve manual motor dexterity. The implementation was validated with the development of a system that, based on the architecture, can dynamically exchange tracking devices. In this case, this validation was used with two low-cost tracking devices: Leap Motion and a prototype of a haptic glove.

This work emphasizes the use of low-cost devices to demonstrate their applicability from a budgetary point of view, considering that most of the explored systems use expensive and high-priced materials, causing them not to be accessible to the public, for this reason, it proposes the use of Leap Motion and which according to a previous study [20, 21] has shown acceptance, accuracy, and effectiveness to capture specific movements for physical therapies.

The rest of the paper is organized as follows: Section II describes the background Section III presents the dynamic architecture proposal and the dynamic architecture applied in

the virtual rehabilitation system, Section IV shows the results and discussion and, Section V describes the conclusions.

## II. BACKGROUND

### A. Related Works

The design of the architecture in virtual rehabilitation systems is a fundamental part of their creation, however, the literature does not show many related works that focus on this point. In Avola's work [11], a framework for the rapid prototyping of virtual rehabilitation systems is proposed. This proposal also includes a description of libraries to be used as well as information on software development such as avatars and stage design. Another proposal by Avola [22] shows a framework to develop low-cost serious 3D games, this system allows physiotherapists to create personalized rehabilitation exercises without special programming knowledge.

Regarding the design of dynamic software architectures, the work of Cuesta [23] informally elaborates a reflexive model of software architectural description with the scope to dynamic systems but does not go into depth in a practical aspect. Other works such as the ones done by León [5] and Ortiza [6] do not focus their proposal on architecture and show proposals for systems with isolated architectures in their case of application.

### B. Software Architecture

The software architecture of a computer program or system is the structure or structures of the system, comprising software components, the externally visible properties of those components, and the relationships between them [24].

Software architecture is important for a wide variety of technical and non-technical reasons. Some of the reasons are: 1) An architecture will inhibit or enable quality attributes, 2) Decisions that are made in an architecture allow reasoning and managing changes as the system evolves, 3) The analysis of an architecture allows an early prediction of the qualities of a system 4) A documented architecture improves communication between interested parts. 5) Architecture defines a set of constraints on the subsequent implementation. 6) Architecture is the key artifact that allows the architect and project manager to reason about cost and schedule. 7) Architecture-based development focuses attention on component assembly, rather than simply creating it [24].

Architecturally significant requirements (ASR) are those requirements that have a significant impact on the architecture of a software system. They are a subcategory of general software requirements.

### C. Dynamic Architecture

The objective of Software Architecture is to describe the structure of systems, and this should include both its static and changing –dynamic– parts. In fact, in many systems, the differential feature, which distinguishes it from other similar systems, is the dynamic behavior of its architecture; often, it can be even more important than the static schema of departure [15].

Systems have grown in diversity and complexity, which is why the need for dynamic architectures now arises. A dynamic

architecture has as its main characteristic the addition, elimination, or replacement of components, sometimes even as part of normal operation, so structurally they must be taken into account during the design [23]. When the evolution of architecture is continuous, and the changes inside it follow a predefined pattern, then it must be considered that this pattern is an intrinsic part of the structure.

## III. METHOD

### A. Dynamic Architecture Proposal

For the dynamic architecture proposal, the steps shown in Table I have been carried out.

TABLE I. DEVELOPMENT METHODOLOGY

Description Methodology
Step 1: Identification of system quality attribute requirements.
Step 2: Identification of architecturally significant requirements.
Step 3: Identification of architecture components.
Step 4: Classification of components: Identification of static and/or dynamic components.
Step 5: Design of architecture components.
Step 6: Documentation of architecture and validation of design decisions.
Step 7: Analysis and evaluation of software architecture.

### B. Dynamic Software Architecture for Virtual Rehabilitation Systems

For architecture design, the steps described in Table I were considered.

1) *Identification of system quality attribute requirements:* Virtual Reality (VR) technology applied to rehabilitation finds its foundation in the field of motor learning [25]. Virtual Reality attributes that align with the motor learning variables (VAM) defined by Levac & Svestrup [26] are practice: enriched environment and motivation (Table II). These variables are related to software quality attributes for systems.

2) *Identification of architecturally significant requirements:* The functional requirements that have been considered to develop the architecture are shown in Table III.

3) *Identification of architecture components:* The approach is presented in four components that are described below

a) *Component a - data capture:* This component seeks to integrate tracking devices to achieve the detection of rehabilitation movements. A sub-component to consider will be a calibration section, to alleviate lighting, latency, and proximity range issues.

b) *Component b - controller:* This component is the one that receives and stores information as established in the configurations of the rehabilitation sessions or movements. A subcomponent to consider will be a data reporter along with its respective connection to the database.

TABLE II. ATTRIBUTES

VAM	Attributes	Software quality attributes
Practice: quantity, task specificity, and meaning	Potential for abundant repetition of practice tests. Ecologically valid VE improves task specificity. Train movements that are identical to those required in real-life tasks. Options to individualize at different challenge levels. Enriched environment. Goal-oriented tasks. Familiarity with commercially available virtual reality gaming systems.	Reliability Usage facility
Enriched environment	Accurate and consistent. Auditory, visual, or tactile. Performance knowledge. Results knowledge. Positive motivational feedback. Motor learning variable.	Efficiency
Motivation	Novelty of VR technology. Game features. Feedback. Goal-oriented tasks. Ability to individualize treatment options. Users can select tasks. Competition with other players. The coincidence between cognitive and physical effort.	Performance

TABLE III. FUNCTIONAL REQUIREMENT

Functional requirements	Architecture module
Hand-object interaction for better presentation.	System: Game for motor manual rehabilitation
Data capture through hardware for visualization and/or necessary calculations for the software	Hardware device: - Optical tracking device - Haptic tracking device
Information saving from hardware (loose data) and/or final results based on the information given by the hardware.	Relational Database

c) *Component c - movement classifier*: This component goes along with movement and gesture recognition functions. It is used to identify the precision with which a patient performs an action.

d) *Component d - virtual scene*: This component offers an interactive and playful rehabilitation environment developed or accepted by medical specialists. Some specific subcomponents will have to manage the duration of the sessions, the number of elements, and control parameters (right or left hand).

4) *Classification of component - Identification of static and/or dynamic components*: Given that there are a variety of options for tracking devices on the market, as well as games or virtual environments, components A and D (Fig. 1) are proposed as dynamic so that, depending on the demographic, budgetary or patient conditions, they can be replaced. or complemented (added). On the other hand, components B and C (Fig. 1) are considered static since they must be the control axis of the system.

5) *Design of architecture components*: Fig. 1 shows the generic design of the architecture proposed together with the four identified components. An application instance of the proposed architecture for each of the components is detailed below.

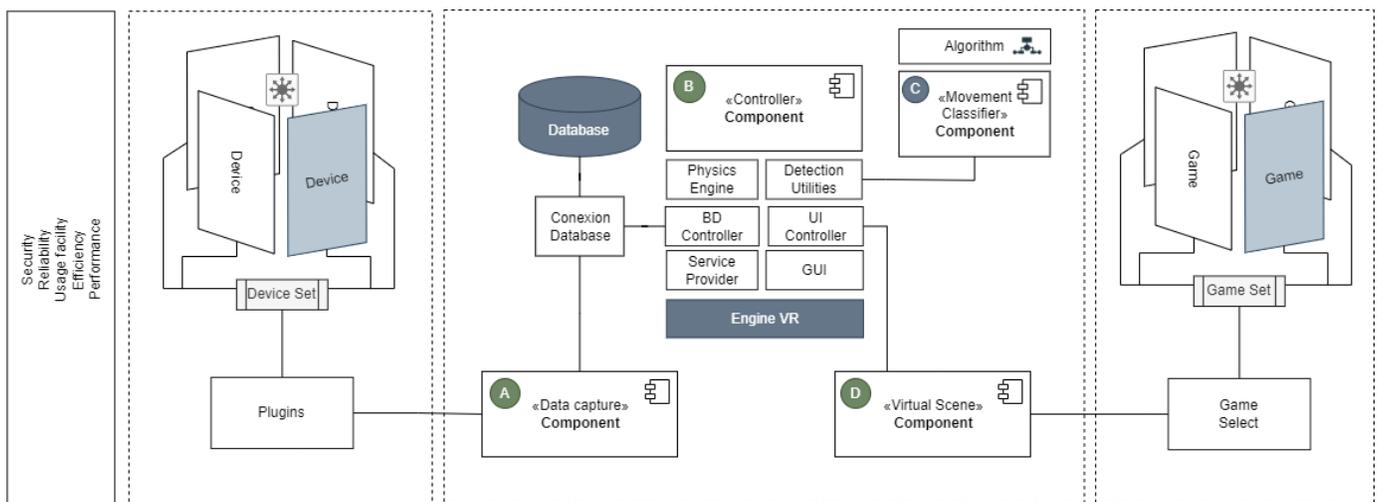


Fig. 1. Generic design of a dynamic software architecture for virtual rehabilitation systems

a) *Component a - data capture*: In this step, two devices have been selected to capture data from the detection of rehabilitation movements: Leap Motion Controller and a Haptic Glove. A calibration section was also considered in this module, to alleviate lighting, latency, and proximity range issues.

Leap Motion Controller (LMC): It is an optical device without markers oriented to control gestural movement. It detects hand movements, fingers, and even objects around its range of vision. LCM can control different interfaces on any computer simply by executing hand movements without touching screens [21]. This device is small and cheap compared to others on the market. LMC dimensions are: 75 mm long, 25 mm wide, and 11 mm high. Its structure consists of two cameras that contain monochrome sensors sensitive to infrared light, whose function is to capture all images [27]. The coverage area (Fig. 2) of the Leap Motion device: it is a hemisphere with a radius of 61 cm, which depends on the viewing angle of the lenses of the two cameras, as well as the maximum intensity provided by the USB connection to the LEDs.

Haptic Glove: The device allows to perform movements in order to control an avatar. This glove uses a sensor that detects supination and pronation movements of the hand, as well as abduction and adduction movements of the wrist.

b) *Component b - controller*: This component must maintain a direct relationship with the data generated in the tracking device and its representation in virtual environments. For this purpose UNITY was selected (Fig. 3)

c) *Component c - movement classifier*: In this work the Dynamic Time Warping (DTW) algorithm is used for sequence recognition [28] to evaluate the degree of similarity of the movements with other previously recorded, given its compatibility with Leap Motion SDK.

d) *Component d - virtual scenario*: Finally, in Component D, a rehabilitation environment was designed through a 2D game. Which focuses on controlling the execution of movements in the hands. This design was based on the International Classification of Functioning, Disability, and Health (ICF) [29] of the WHO and focused on exercises that make it possible to enhance the component of mobility-oriented activities to improve motor dexterity.

The specific activities to be implemented, according to the ICF classification, refer to subclassification d4453: "Turn or twist hands or arms" and subclassification d4400: Pick up:

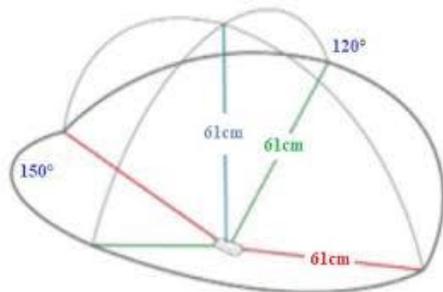


Fig. 2. Leap motion coverage area

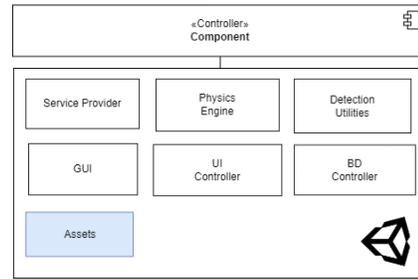


Fig. 3. Controller subcomponents

"lift or take a small object with hands and fingers, as when picking up a pencil. These exercises were proposed by two physical therapy specialists and a pediatric neurologist. The implemented gestures were extension, flexion, supination, and pronation, and the combination of them to achieve displacement and jumps. The tracking device (Leap Motion or Haptic Glove) allows to control the game avatar.

The objective of the game is to mobilize an avatar through a road which must avoid a certain number of obstacles through jumps until reaching a destination. Every time it avoids an obstacle it will receive points and/or prizes depending on the difficulty, and every time it misses it loses one point.

The game has three difficulty levels which refer to the proximity of the obstacles and the reaction time of the patient to jump. Fig. 4 shows avatar control using Leap Motion and Fig. 5 shows avatar control with the haptic glove.

Fig. 6 shows the running game environment. Since many types of patients can be trained in motor rehabilitation sessions, the system must be adapted to the different needs of each patient.



Fig. 4. Game in execution, controlled with leap motion



Fig. 5. Game in execution, controlled with the haptic glove

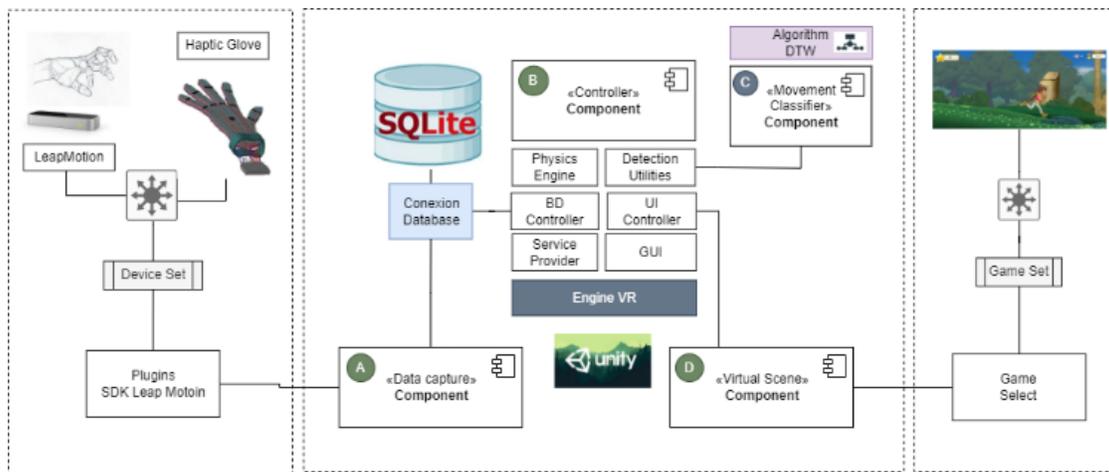


Fig. 6. Dynamic software architecture for a virtual rehabilitation system for the improvement of manual motor dexterity

The game allows different configurations according to the characteristics of the patient: Sex (Male or Female), Age (allows to use a different avatar). The game allows to set the time for the exercises, the number of obstacles, and which hand to use (left or right hand). These features are a clear example of the need for dynamic architecture.

6) *Documentation of architecture and validation of design decisions:* As a result of the application of the architecture, the specific implementation of the architecture is shown in Fig. 7. The validation of design decisions is very complex and generally depends on the context. For validation, there are qualitative techniques such as scenarios, questionnaires, or checklists and quantitative techniques with metrics, simulations, mathematical models, prototypes, or experiments [30].

It was determined to validate the design by a quantitative technique performing a simulation of rehabilitation, experimentation, and a specific analysis because of the changing nature of the system. This point is described in detail in step 7.

7) *Analysis and evaluation of software architecture:* For the evaluation, a comparative analysis of the user experience (UX) was used, considering the implementation with the LeapMotion optical tracking device and the implementation with the haptic glove. The instrument was the validated User Experience Questionnaire (UEQ) [31].



Fig. 7. Game in execution, shows the control of time to achieve the goal by avoiding the obstacles. points and reward system is managed

UEQ comprehensively measures six dimensions: (a) Attractive: the product must look attractive, pleasant, and friendly; (b) Efficiency: the user must perform tasks with the product quickly, efficiently, and pragmatically; (c) Perspicuity: the product must be easy to understand, clear, simple and easy to use; (d) Reliability: the interaction with the product must be predictable, safe and meet the expectations; (e) Stimulation: the use of the product must be interesting, exciting and motivating; (f) Novelty: the product must be innovative, inventive and creatively designed [32].

UEQ consists of 26 questions and applies a seven-point Likert scale. The questionnaire is shown in Fig. 8.

The population size was university students from a public university. The sample was non-probabilistic for convenience. There were participants from schools of Medicine, Psychology, Engineering, and Natural Sciences of Education and Accounting. The age range was between 18 and 22 years old, there were 50 men and 56 women with a total of 106 participants.

	1	2	3	4	5	6	7		
desagradable	<input type="radio"/>	agradable	1						
no entendible	<input type="radio"/>	entendible	2						
creativo	<input type="radio"/>	sin imaginación	3						
fácil de aprender	<input type="radio"/>	difícil de aprender	4						
valioso	<input type="radio"/>	de poco valor	5						
aburrido	<input type="radio"/>	emocionante	6						
no interesante	<input type="radio"/>	interesante	7						
impredecible	<input type="radio"/>	predecible	8						
rápido	<input type="radio"/>	lento	9						
original	<input type="radio"/>	convencional	10						
obstruivo	<input type="radio"/>	impulsor de apoyo	11						
bueno	<input type="radio"/>	malo	12						
complicado	<input type="radio"/>	fácil	13						
repeler	<input type="radio"/>	atraer	14						
convencional	<input type="radio"/>	novedoso	15						
incómodo	<input type="radio"/>	cómodo	16						
seguro	<input type="radio"/>	inseguro	17						
activante	<input type="radio"/>	adomecedor	18						
cube expectativas	<input type="radio"/>	no cube expectativas	19						
ineficiente	<input type="radio"/>	eficiente	20						
claro	<input type="radio"/>	confuso	21						
no pragmático	<input type="radio"/>	pragmático	22						
ordenado	<input type="radio"/>	sobrecargado	23						
atractivo	<input type="radio"/>	feo	24						
simpático	<input type="radio"/>	antipático	25						
conservador	<input type="radio"/>	innovador	26						

Fig. 8. UEQ questionnaire

During the evaluation, each participant receives an explanation of the use of the system. Then each participant used the system with their dominant hand. Each participant had a 20-minute interaction with the system, 10 minutes with the Leap Motion device, and 10 minutes with the haptic glove.

IV. RESULTS AND DISCUSSION

A. Results

Table IV shows the results for the use of the Leap Motion Controller device. Values between -0.8 and 0.8 represent a more or less neutral evaluation of the corresponding scale, values > 0.8 represent a positive evaluation, and values < -0.8 represent a negative evaluation. The range of the scales is between -3 (terribly bad) and +3 (extremely good) [33].

As shown in Table IV and Fig. 9, all the values are positive for the use of Leap Motion, with Perspicuity being the highest value, followed by the attractiveness value. For our study and the evaluation of the architecture, the most relevant values are related to the efficiency and reliability scales. These values are positive according to the scale but could be improved.

As shown in Table V and Fig. 10, the evaluation of the use of the system with the Haptic Glove, positive values have also been obtained within the Evaluation Scale, but a little lower ones compared to the use of Leap Motion. Focusing on the Efficiency and Reliability values, both are very similar to the values obtained with Leap Motion.

TABLE IV. RESULTS OF LEAP MOTION EVALUATION

UEQ OF LEAP MOTION EVALUATION		
Attractiveness	1.987	0.96
Perspicuity	2.068	1.31
Efficiency	1.517	1.15
Dependability	1.649	0.95
Stimulation	1.703	1.16
Novelty	1.401	1.33

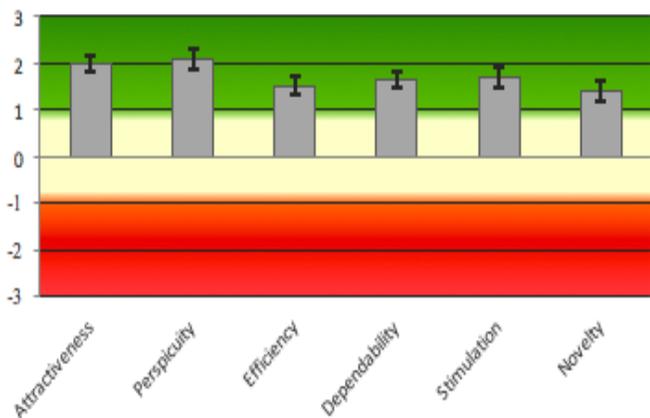


Fig. 9. Leap motion results

TABLE V. RESULTS OF HAPTIC GLOVE EVALUATION

UEQ OF LEAP MOTION EVALUATION		
Attractiveness	1.930	0.90
Perspicuity	1.988	1.17
Efficiency	1.517	1.15
Dependability	1.623	0.93
Stimulation	1.696	1.18
Novelty	1.380	1.29

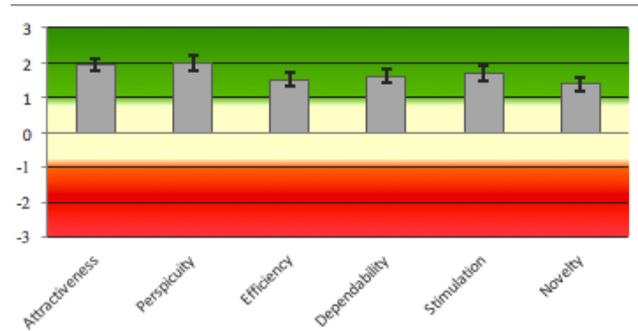


Fig. 10. Results of haptic glove

B. Discussion

According to our User Experience results, the proposed software architecture gave the system through the non-functional requirements of efficiency and reliability enough dynamism so that users not only use different tracking devices for virtual rehabilitation but also in regarding the ease of use, the preliminary users did not feel a change in terms of performing the tasks with the device. Added to this is the fact of performance, since none of the participants reported discomfort and all described the system as a tool support for active participation during the rehabilitation process.

This work provides a contribution by focusing on the architectural aspect of this type of system, since unlike other works where solutions to mobility and rehabilitation problems are proposed, this area is not deepened, leaving said design proposals in isolated environments, traditional and poorly approaches documented.

In works such as [6-8], architectures such as session management and feedback modules are proposed to provide information to therapists or doctors and patients respectively. It is considered that the administrative aspect is not the focus of this proposal but it can be implemented through the integration of the UI and database controller subcomponents. In addition, it will be up to the developer to decide how to complement each exposed component according to its context and rehabilitation tools.

This work does not cover the communication complexities as it works [7] that are usually manifested in customized constructions of tracking devices, where serial or wireless bridge models are used, this prototype uses one of these devices, however, these problems are solved by moment of implementation with the use of external libraries with support for Unity.

It is emphasized that in all cases it is mentioned that the interfaces must be designed with the aim of guiding a high usability which is an aspect that is dealt with in this work.

The architecture prototype was validated with low-cost devices, which demonstrates its versatility so that it can be implemented for systems that have been shown and seek effectiveness [34] to function in any clinical or home environment. This aspect is also reinforced by showing that the devices can be selected from a broader universe that can provide a variety of alternatives regarding costs or patient comfort to treat much less invasive or intrusive devices [35] and adapt to their needs.

This work also proposes to make the rehabilitation environment or game dynamic, in the prototype it was implemented with a single option to make it similar to traditional systems and guide user tests in a single line, however it is recommended to expand the application in groups separated from users, of more varied ages and longer usage times, since such dynamism is easier to achieve once the tracking device is integrated. For future lines of research, it may also be advisable to increase the sample size and include a control group to check the effectiveness of the system with users with manual motor problems. Our results provide evidence of the application of the architecture in previous environments with healthy users.

## V. CONCLUSION

This work showed the design of a dynamic software architecture for low-cost virtual rehabilitation systems for the improvement of manual motor dexterity. The development had the constant advice of medical specialists in physical therapy who identified the exercises to be performed: pronation/supination and flexion/extension for the application instance. The proposed architecture has two static components for the administrative control of the system and two dynamic components to integrate gaming environments and tracking devices. This type of architecture is proposed to delve into this type of system and adapt to the various diseases that cause problems in motor dexterity.

The proposed architecture shows the feasibility through its application instance of reusing and integrating the various tracking devices as well as the different game interfaces that have been proposed in the literature in this gap.

The development considered the required attributes of virtual rehabilitation systems for therapies. With respect to other immersive-type systems, Leap Motion has the advantage that it does not cause discomfort in terms of dizziness and/or nausea, in addition to its characteristics that there is no physical contact with the skin, which is an advantage when patients are children.

The limitations of the study are given by the time of use of each device. As future work, it is intended to implement the suggestions received by the participants, and to carry out the validation of the architecture with other data capture devices and games.

## ACKNOWLEDGMENT

This contribution was financed by the “Universidad Nacional de San Agustín de Arequipa” under contract IB-42-2020-UNSA project “Virtual Rehabilitation System (VR) for motor and cognitive improvement in children with Epileptic Encephalopathy. CEPIVIRT”.

## REFERENCES

- [1] M. Alenezi, Software Architecture Quality Measurement Stability and Understandability, *International Journal of Advanced Computer Science and Applications*, vol. 7, no. 7, 2016.
- [2] L. Fernandez. CIMAT Metodologías ágiles - Arquitectura de Software, vol. 02, no. 03, 2006.
- [3] R. M. Patterson and V. Priganc, “What does hand rehabilitation look like in 50 years?”, *Journal of Hand Therapy*, vol. 33, no. 2, pp. 269–270, Apr. 2020.
- [4] E. Monge Pereira, F. Molina Rueda, I. M. Alguacil Diego, R. Cano de la Cuerda, A. de Mauro, and J. C. Miangolarra Page, “Empleo de sistemas de realidad virtual como método de propiocepción en parálisis cerebral: guía de práctica clínica”, *Neurología*, vol. 29, no. 9, pp. 550–559, Nov. 2014.
- [5] J León, “Software de Realidad Virtual para rehabilitación de enfermedades neuromusculares”, Escuela Colombiana de Ingeniería Julio Garavito, 2014. <https://repositorio.escuelaing.edu.co/handle/001/196>
- [6] M. Ortiz, “InTrainer Sistema de rehabilitación cardiaca aumentado por realidad virtual”, Universidad del País Vasco / Euskal Herriko Unibertsitatea, Departamento de Ciencia de la Computación e Inteligencia Artificial, San Sebastián, 2010.
- [7] J. Solana, O. Renda, C. Cáceres, P. Rodríguez, A. Serrano, E. Opisso Salleras, P. Cingolani, P. Giorgianni, R. Sánchez, J.M. Tormos, S. Scattareggia and E.J. Gómez, Arquitectura interoperable de tele-rehabilitación domiciliaria.
- [8] F. Moreno, J. Ojeda, E. Ramírez, C. Mena, O. Rodríguez. Un Framework para la Rehabilitación Física en Miembros Superiores con Realidad Virtual, Centro de Computación Gráfica, Escuela de Computación. Facultad de Ciencias, Universidad Central de Venezuela, Caracas, Venezuela.
- [9] Gesture Therapy. Available: [http://robotic.inaoep.mx/~foe/blog/?page\\_id=6](http://robotic.inaoep.mx/~foe/blog/?page_id=6)
- [10] V. Tundjungsari, A. S. M. Sofro, H. Yugaswara, and A. T. D. Putra, “Development of mobile health application for cardiovascular disease prevention”, *International Journal of Advanced Computer Science and Applications*, vol. 9, no. 11, 2018.
- [11] D. Avola, M. Spezialetti and G. Placidi, “Design of an efficient framework for fast prototyping of customized human-computer interfaces and virtual environments for rehabilitation”, *Computer Methods and Programs in Biomedicine*, vol. 110, no. 3, pp. 490-502, 2013.
- [12] G. Vivas Mateos, “Desarrollo de un método de captación de movimiento humano para el control remoto de terapias de rehabilitación robóticas”, final degree project, Polytechnic University of Madrid, Madrid, España, 2016.
- [13] I. Dimbwadyo-Terrer, F. Trincado-Alonso, A. de los Reyes-Guzmán, M. A. Aznar, C. Alcubilla, S. Pérez-Nombela, A. del Ama-Espinosa, B. Polonio-López and Á. Gil-Agudo, “Upper limb rehabilitation after spinal cord injury: a treatment based on a data glove and an immersive virtual reality environment”, *Disability and Rehabilitation: Assistive Technology*, vol. 11, no. 6, pp. 462-467, 2016.
- [14] R. J. Adams, A. L. Ellington, K. Armstead, K. Sheffield, J. T. Patrie and P. T. Diamond, “Upper extremity function assessment using a glove orthosis and virtual reality system”, *OTJR: Occupation, Participation and Health*, vol. 39, no. 2, pp. 81-89, 2019.
- [15] A. Navasa, “Marco de trabajo para el desarrollo de Arquitecturas de Software”, Universidad de Extremadura , pp. 81- 84, 2008.

- [16] K. C. Nickels and E. C. Wirrell, "Cognitive and Social Outcomes of Epileptic Encephalopathies", *Seminars in Pediatric Neurology*, vol. 24, no. 4, pp. 264–275, 2017.
- [17] R. Teasell *et al.*, "A systematic review of the rehabilitation of moderate to severe acquired brain injuries", *Brain Injury*, vol. 21, no. 2, pp. 107–112, Feb 2007.
- [18] D. De Noreña, M. Ríos-Lago, I. Bombín-González, I. Sánchez-Cubillo, A. García-Molina and J. Tirapu-Ustárroz, "Efectividad de la rehabilitación neuropsicológica en el daño cerebral adquirido (I): atención, velocidad de procesamiento, memoria y lenguaje", *Rev Neurol*, vol. 51, no.11, pp. 687-98, 2010.
- [19] World Health Organization, <https://www.who.int/>
- [20] E. E. Saavedra Parisaca and E. Enriqueta Vidal Duarte, "Low-Cost Optical Tracking Controller System for Fine Motor Rehabilitation in Children with Brain Damage: Formal Specification and Validation", *16th Iberian Conference on Information Systems and Technologies (CISTI)*, Jun. 2021.
- [21] P. Wozniak, O. Vauderwange, A. Mandal, N. Javahiraly, and D. Curticapean, "Possible applications of the LEAP motion controller for more interactive simulated experiments in augmented or virtual reality", *SPIE*, vol. 9946, pp. 234-245, September 2016.
- [22] D. Avola, L. Cinque, G. L. Foresti and M. R. Marini, "An interactive and low-cost full body rehabilitation framework based on 3D immersive serious games", *Journal of biomedical informatics*, vol. 89, pp. 81-100, 2019.
- [23] C. Cuesta, "Arquitectura de software dinámica basada en reflexión", Biblioteca Virtual Miguel de Cervantes, 2022. Available: <https://www.cervantesvirtual.com/obra/arquitectura-de-software-dinamica-basada-en-reflexion--0/>
- [24] L. Bass, P. Clements and R. Kazman, *Software Architecture in Practice*. Addison Wesley, 1999, ISBN 0-201-19930-0.
- [25] M. K. Holden, "Virtual environments for motor rehabilitation: Review", *Cyberpsychology and Behavior*, vol. 8, no. 3, pp. 187–211, 2005.
- [26] Danielle E. Levac and Heidi Sveistrup, "Virtual Reality for Physical and Motor Rehabilitation", Chapter 3, *Motor Learning and Virtual Reality*, 2014.
- [27] F. Weichert, D. Bachmann, B. Rudak, and D. Fisseler, "Analysis of the Accuracy and Robustness of the Leap Motion Controller", *Sensors*, vol. 13, no. 5, pp. 6380–6393, May 2013.
- [28] P. Senin, "Dynamic time warping algorithm review", Information and Computer Science Department University of Hawaii at Manoa Honolulu, USA, vol. 855. pp. 1-23, 2008.
- [29] International Classification of Functioning, Disability and Health, 2021 Available: <https://www.who.int/standards/classifications/international-classification-of-functioning-disability-and-health>.
- [30] O. Salvador, "Evaluando la Arquitectura de Software: Parte 1. Panorama General", SG Buzz, 2022, Available: <https://sg.com.mx/content/view/235>.
- [31] User Experience Questionnaire, Available: <https://www.ueq-online.org/>
- [32] A. Al-Hunaiyyan, R. Alhajri, B. Alghannam, and A. Al-Shaher, "Student Information System: Investigating User Experience (UX)", *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 2, 2021.
- [33] B. Laugwitz, T. Held, and M. Schrepp, "Construction and Evaluation of a User Experience Questionnaire", *Lecture Notes in Computer Science*, pp. 63–76, 2008.
- [34] C. Meneses Castaño, P. Penagos, and B. Yamile Jaramillo, "Efectividad de la tecnología robótica y la realidad virtual para la rehabilitación de la función motora en la parálisis cerebral. Revisión sistemática", *Rehabilitación*, pp. 100752, Nov. 2022.
- [35] F. Manuel, "Diseño de un sistema basado en tecnologías móviles, wearables y análisis de datos para promover el envejecimiento activo en mayores", *Digibug.ugr.es*, Dec. 2022.

# Insights into Search Engine Optimization using Natural Language Processing and Machine Learning

Vinutha M S<sup>1</sup>, M C Padma<sup>2</sup>

Research Scholar, Department of Computer Science & Engineering, P E T Research Centre, Mandya,  
University of Mysore, Mysuru, India<sup>1</sup>

Professor, Department of Computer Science & Engineering, PESCE, Mandya, India<sup>2</sup>

**Abstract**—Among the potential tools in digital marketing, Search Engine Optimization (SEO) facilitates the use of appropriate data by providing appropriate results according to the search priority of the user. Various research-based approaches have been developed to improve the optimization performance of search engines over the past decade; however, it is still unclear what the strengths and weaknesses of these methods are. As a result of the increased proliferation of Machine Learning (ML) and Natural Language Processing (NLP) in complex content management, there is potential to achieve successful SEO results. Therefore, the purpose of this paper is to contribute towards performing an exhaustive study on the respective NLP and ML methodologies to explore their strengths and weaknesses. Additionally, the paper highlights distinct learning outcomes and a specific research gap intended to assist future research work with a guideline necessary for optimizing search engine performance.

**Keywords**—Search engine optimization; google search; natural language processing; machine learning; recommendation

## I. INTRODUCTION

In the present era of the competitive market, every organization and individual intends to ensure that their information reaches the right clients in minimal effort. Stakeholders also need to have a clear insight into their upcoming business demands. From all these perspectives, business products and services are usually maintained via websites. This target is met by using Search Engine Optimization (SEO), which facilitates carrying out the operation to assist the client webpage or its contents to offer higher ranks on the standard platform of Google [1]. The prime distinction between paid advertisement and SEO is that SEO uses an organic methodology to generate ranking scores [2][3]. It will eventually mean that a user will not be required to pay to be in that environment of using SEO [4]. In simplified form, the user of SEO tools identifies and extracts suitable content from a target webpage and optimizes it so that the webpage always appears at the top of google searches [5]. SEO tools, therefore, assist in making higher visibility of the webpage and offers a higher probability of reaching the maximum number of customers. There standard operational taxonomy of SEO is of two types, i.e., on-page and off-page process SEO [6][7]. Basically, the ranking associated with the webpage can be improved by appropriately building the web content using an on-page process of SEO. This process essentially includes constructing higher-quality content, generating appropriate keywords, managing meta-tags, and enhancing the different

objects to ensure that it is well-chosen by the target customer in the on-page SEO process. On the other hand, the backlinks' optimization process is carried out at the backend of the webpage in the off-page process of SEO. This form of SEO mainly focuses on establishing relationships among the content to reach its appropriate customer. Currently, a specific set of programs called bots are used to perform crawling within the webpage using existing search engines, viz. Bing/Google [8]. This operation aggregates information associated with the target web contents, placing them in the form of an index. The web contents are analyzed within the index by such algorithms considering a massive number of signals or ranking values. This is done to ensure the availability of the page at the top of query hits. The prime target of such a form of the search algorithm is to evolve up with a highly authoritative page to offer a superior experience of searching by the user. Irrespective of all the efforts towards improving the performance of SEO tools, there are still serious concerns that have posed as an impediment, viz. i) inaccurate formulation of webpage index, ii) identifying and constructing a precise keyword, iii) structuring the wrong webpage/contents not in line with the target topic, iv) internal linking to be highly incoherent, v) slower /fluctuating uploading performance of the web page in different computing device [9][10]. Therefore, this paper identifies the potential of using Natural Language Processing (NLP) and Machine Learning (ML) approaches to improve the performance of SEO. The paper contributes to potential learning outcomes from existing literature. Further, it also contributes towards identifying significant research gaps extracted from existing techniques to improve SEO performance.

The paper's organization is as follows: Section II discusses the fundamental information about SEO, followed by reviewing existing research practices of SEO with NLP in Section III. Section IV discusses ML practices used in SEO while Section V discusses existing SEO tools. A discussion of existing research trends of SEO is carried out in Section VI, while the research gap is highlighted in Section VII. Section VIII makes discussion about the results and research implications. Section IX finally concludes the paper with significant learning outcomes followed by a briefing of future work to be carried out.

## II. INSIGHTS ON SEO

This section presents insights into SEO or Website positioning. Firstly, a brief description of SEO followed by a

working principle of a search engine is discussed to understand the intrinsic mechanism of SEO. Further, factors affecting website posting are briefly discussed, and finally, this section discusses challenges in SEO.

### A. Search Engine Optimization

SEO is an act that includes a series of professional activities. These activities include the practice of improving the structure of content, thereby increasing visibility in search engines and gaining a large amount of traffic to the website [11]. Common SEO practices include rich content creation, keyword optimization, and link building. Thus, SEO is a powerful mechanism for advancing search engine algorithms to come up with the most relevant and appropriate web content and improve the website's ranking (in an organic way) in search results, ultimately boosting marketability and increasing sales.

### B. Work Principle of the Search Engine

Search engines are obviously fundamental to the SEO process, but many practitioners are unaware of how they work. Therefore, one must first comprehend the basics of search engines to learn SEO. A search engine is a service that enables web search by performing three important tasks such as crawling, indexing, ranking, and recognizing items in the system record or database corresponding to keywords specified by the user [12-13]. Crawling enables search engines to discover content, and indexing is a mechanism for obtaining web documents and maintaining replicas of the content they have visited. The ranking is mainly subject to search engines mainly concerned with SEO operations. Fig. 1 depicts the schematic architecture of the web content indexing process.

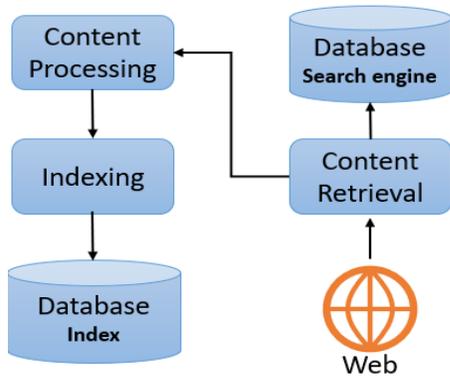


Fig. 1. Procedure of web content indexing

Web content is retrieved using a WebCrawler (bot) that stores the web content in a database of search engines. In addition, web content is subject to data processing operations such as stemming, HTML tag, and stop-word removal. Later, indexing is done by search engines by generating direct and replicating content, such as single words and their positional information on the search page. Furthermore, the search engine keeps the indexes in its index database. Fig. 2, depicts the schematic architecture of the content querying and retrieval procedure.

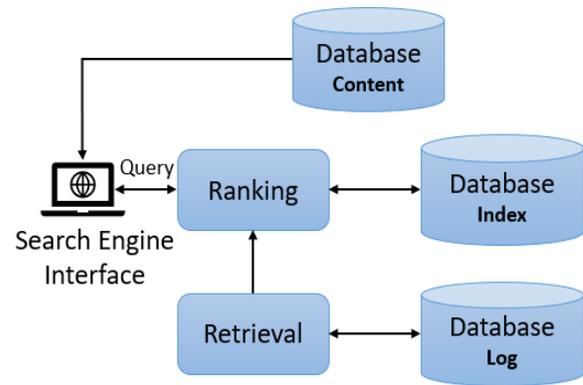


Fig. 2. Content querying and retrieval

Through the search engine interface, the user provides a search query. The search engine algorithm creates a URL ranking list that matches the user's query to the index database based on contextual information. The search engine then displays a snippet subjected to the ranked URL to the user, who can browse and select to retrieve the corresponding content in its original form from the content database.

### C. Factors Affecting Ranking and SEO Challenges

The ranking of web content is influenced by many factors, including page relevance, temporal factors, and link weights [14]. A webpage's relevance is determined by its tags, density distribution, and identical keywords. Temporal aspects are concerned with the oldness of websites, web contents and webpages, the oldness of links, and the duration of domain registration. There are both internal and external links in the contents. However, the external link is given more weight as it is associated with significant factors such as quality, quantity, relevancy, and repetition. A basic mechanism of SEO includes almost all the core attributes of the above-discussed factors, which can be numerically simplified and expressed as follows:

$$Seo = \int C + L + K + O \quad (1)$$

Where,  $C$  is the web content,  $L$  denotes link,  $K$  refers to user keyword,  $O$  represents other factors such as oldness of website, or blog, server, web-design, URL, domain name, and many more. All these factors have priority and should follow priority order as mentioned in the above expression 1. Apart from this, a few challenges associated with search engines significantly affect the quality of the SEO process [15-16]. The first major issue is content spamming, a common method used by unethical users to get their web pages in top results. The next issue is the article spinning, similar to scraping data using specialized software that takes the copied original and reproduces it as a new, original article for future use. The third issue is keyword stuffing, in which users reuse keywords like name, meta, head, etc., in different HTML tags and URL spammers. Furthermore, masquerading is an SEO technique used to mislead users by redirecting them to a page that is different from the page crawled by search engines. Similarly, a URL redirection is also a significant issue where the file is redirected to a specific URL as soon as the user loads the site.

### III. REVIEW OF SEO APPROACHES USING NLP

NLP is an area of ML that reveals the precise structure and meaning of content. Modern websites are driven by algorithms, which determine what they display in search results for specific keywords. Using NLP in optimizing web content, it can be expected that the content would reach the top of the search rankings. NLP can be used to analyze website content and optimize it for specific keywords or phrases. It can be used to identify and correct grammar and spelling errors, as well as to generate content that is optimized for search engines. NLP techniques can also be used to analyze user queries and optimize website content to better match those queries

Many research works are using the mechanism of NLP to achieve optimization in the search ranking. This section provides a brief highlight of the existing literature in the context of SEO. A research article presented by Killoran et al. [17] has examined the influential factors that have a high impact on search ranking. It is reported that search ranking is formed on the basis of participants' category, SEO experts, search engine companies, and users. During the choice of keywords, the authors stated that the website's target audience and competitors have to be taken into account. The study concludes that a combination of appropriate keyword placement and link-building may yield the desired solution. The study of Hajeer et al. [18], applied the NLP mechanism to overcome the limitations associated with the Porter algorithm used for term normalization and index time reduction in the content retrieval systems. The authors have presented a different stemming technique to enhance content searching in an information retrieval system. The results claim improvement over existing technologies. Tsuei et al. [19] devised a customized decision model based on the interview and survey for SEO in internet marketing to boost the hit rate of websites on the search page that satisfy users' requirements. The finding of this study suggests that meta tags are the most influential factor that has a significant impact on the search ranking.

The work of Luh et al. [20] aimed to examine the ranking mechanism of the Google search engines from an SEO viewpoint. The study suggested an estimation function for determining the score of query matching from a limited set of ranking factors. Further, re-ranking is carried out on the basis of obtained scores. The scope of the presented scheme is evaluated based on the comparison of newly obtained ranks with the original ranks. Jenkins et al. [21] developed a model for constructing text annotations for SEO. This model employs the Extreme Gradient Boosting algorithm for precise labeling phrases. Also, logistic regression is considered in this model to generalize the rank of aggregated annotations for clusters of content. The study findings demonstrate that the presented model increases the traffic to the web content by 1-2%. A semantic architecture using web and data mining techniques is presented by Sharma et al. [22] for personalizing the eCommerce search engine. The design and development of the architecture consist of a series of implementation phases were, firstly, a query expansion is performed to transform the input user query using NLP operations to understand the user

requirement. Afterward, ontology classification is carried out to filter out the relevant subjects of the web content. Further topic modeling is carried out using clustering, and statistical computation is then carried to perform a re-ranking operation. Semantic annotation for semi-structured data on a web page using header identification and object classification is presented by Zhang et al. [23]. The authors have designed a description framework for annotating the data domain, and header identification is carried for annotating data objects on the webpage. In addition, a feature vector is constructed for data objects which are left by header identification, and a neural network is then applied to perform semantic annotation.

Adoption of the latent semantic analysis for SEO is carried out by Horasan [24]. In this study, the keyword extraction process from textual data with latent semantic analysis is performed to draw a relationship between documents/sentences and terms in the text using linear algebra. Uzun [25] suggested a model-based string technique and DOM tree for content extraction. The string technique extracts information with the HTML tags followed by the crawling process. The study of Barrett et al. [26] presented an approach for searching large video corpora for clips depicting human language queries expressed as sentences. In this study, a compositional semantics scheme is applied to encode refined meaning to extract the differences between two phrases with the same words under a different context. Sal et al. [27] used a disseminated cooperative cache based on evolutive summary counters to store approximate records of data accesses in a search engine. Ghanbarpour and Naderi [28] examined the ranking technique for keyword search according to the relevancy of the query over graph-structured data. Soltani et al. [29] employed an approach of semantic search engines to develop a different model for software signature search engines. The authors have used the document-to-vector model to compute the signature and user query vectors.

The work of Dai et al. [30] suggested an efficient and adaptive semantic-based keyword ranked search technique using Doc2Vec for secured cloud data. Chen [31] focuses on adopting a user interaction approach to control linguistic ambiguity to improve search engine outcomes. Zhang et al. [32] have suggested a scheme to recognize the identifiers that are associated with semantic text queries. In order to enhance text queries, the authors have looked for keywords within class names from APIs with semantically related APIs. However, if the corpus projects do not have sufficient vocabulary, this technique may not work as well. Calvillo et al. [33] presented an automated mechanism to classify and locate research information based on NLP. The implementation of this scheme focuses on cleaning data by removing aspects such as images and words that are not significant. The digital library was used to extract a percentage of the content from different articles such as abstract, introduction, keywords, and other segments of the article, which help to perform the tests. Hamzei and Hakimpour [34] introduced a method for analyzing queries for spatial search engines. This method employs iterative query segmentation identification of location-names and spatial relationships. Table I highlights the summary of the work being discussed in this section.

TABLE I. SUMMARY OF SEO USING NLP TECHNIQUES

Authors	Problems	Techniques	Advantage	Limitation
Killoran et al. [17]	Search ranking	Analytical study	Highlights influential factors, and important suggestion	Only theoretical and analytical discussion
Hajeer et al. [18],	Indexing time	Stemming	Outperforms Porter algorithm	Related to Over-stemming and only 2.3% of improvement
Tsuei et al. [19]	Identification of factor affecting SEO driver	Decision making system	Highlighted significant website SEO factors	subjective opinions of decision-makers
Luh et al. [20],	Examine ranking of the Google search engines	Rank estimation function and re-ranking scheme	Achieved the best SEO effectiveness	Considered limited set of ranking factors
Jenkins et al. [21]	Understanding content to attract new user	XGBoost and Linear regression	Increases traffic by 1-2 %	Higher dependency on more keystrokes.
Sharma et al. [22]	SEO for ecommerce	Ontology and Semantic Approach	Provides context-aware results and recommendations	Lacks statistical outcome analysis to justify its usability
Zhang et al. [23]	Identification for highlights of multimedia file system	Annotation, Header recognition, Neural network	Semantic annotation of semi-structured information	Only applicable for Chinese language
Horasan [24]	Adoption of knowledge contents	Latent semantic analysis and linear algebra	Complies with the SEO criteria, helpful for who do not know SEO	No effective benchmarking
Uzun [25]	Time efficiency in Web scrapping	String technique and DOM tree	Achieves time efficiency in web scrapping	Dependency on various manual process
Barrett et al. [26]	searching large video corpora from text query	searching large video corpora	Does not require any prior video annotation	Computationally in-efficient
Sal et al. [27]	understanding the underlying content of multimedia	Cooperative cache scheme	Flexible to support large data for analysis	Domain-dependent implementation
Ghanbarpour and Naderi [28]	ranking search problem	Model-based ranking function	Improves the accuracy of the ranking	Only support single keyword search
Soltani et al. [29]	Digital security	Paragraph Vector Model	Achieves higher recall rate	Computationally expensive
Dai et al. [30]	multi-keyword ranking search	Doc2Vec model	Simplified structural model	documents may be lost in the encrypted forms
Chen [31]	misinterprets the user query	Personalized topic search system	Quick response to user search needs	Limited to English language and used small dataset
Zhang et al. [32]	Recognition of the identifiers that are associated with semantic text query	Neural network model (CBOW)	Provides a good scope	If the corpus projects do not have sufficient vocabulary, this technique may not work as well.
Calvillo et al. [33]	locating research paper	NLP based SEO	Better performance in the classification of research article	does not capture position in text,
Hakimpour [34]	analyzing queries for spatial search engines	Iterative query segmentation and spatial relationships.	Better interaction between the users and the search application	Induces to spatial complexity

#### IV. REVIEW OF SEO APPROACHES USING ML

The prime objective of any SEO approach is to find the targeted content which could meet the expectation of the user and thereby make the web content available to them with least effort. This operation demands a better for of optimization, where Machine Learning (ML) approach plays a significant contributory role. ML can help SEO professionals by analyzing the vast amounts of data required to optimize a website's ranking. For instance, it can be utilized to search ranking factors to get insight into website age, bounce rate, and content length. These were significant indicators of high-ranking websites. ML can also help predict future search engine

algorithm changes, enabling SEO professionals to make proactive adjustments. Overall, the use of ML in SEO offers numerous benefits, including increased accuracy in predicting search engine algorithms, automation of SEO tasks, and the ability to analyze large amounts of data. This section briefs about some of the literatures where ML approaches has contributed towards this optimization process considering various forms of use-cases.

The most recent work carried out by Boppana and Sandhya [35] have used Recurrent Neural Network (RNN) in order to facilitate a better form of recommendation system to be used in SEO operation with perspective to web crawling practices. The

core target of this work is mainly to reduce the error while recommending the popularity of extracted information. A clustering approach based on extracted features from contextual information is implemented in this process. The work carried out by Burgess et al. [36] address the problems associated with security of web-contents, which is another essential concern in SEO process. The authors have used Long Short-Term Memory (LSTM) for identifying the possible threat in traffic associated with web-content while making redirection in HTTP. The study claims of successful control of such malicious redirection. Similar aspect of security consideration was also witnessed in investigation carried out by Liu and Fu et al. [37] where an SEO tool is required to confirm the vulnerability in the web-contents. The solution is provided by the author by considering phishing attack on web contents where feature learning is used. The study has used an unsupervised learning methodology in order to identify the insecure web-contents. Further, the model has also used a random walk of biased nature considering fusion of information over URL and structural information.

Soliman et al. [38] have implemented a model using random forest for addressing the need of semantics and linked data of the web-contents. The implementation has used Resource Description Framework (RDF) where random forest is used for retrieving the current state of RDF for assisting in further classification process. Study in the direction of the recommendation system in SEO is also reported in work of Ismail et al. [39], where the focus is mainly towards customizing the recommendation system over web-contents. The study model has used fuzzy logic concept integrated with structural analysis for achieving adaptive recommendation system. Label propagation is another essential target to be achieved in SEO and it becomes quite challenging in presence of heterogenous information. Study in such problem is addressed by Hisano et al. [40] by storing a voluminous information in the form of a network followed by applying Jacobian iteration for learning weights. This technique also contributes towards performing better analysis. It should be noted that web-contents consideration in SEO will also be inclusive of presence of multi-media file systems too. It is found that identification of highlights of such files is completely dependent on trained data curated by human. This hinders scalability as well as is expensive in nature of deployment. This problem is addressed in work of Kim et al. [41] by introducing a ranking mechanism using deep learning technique in presence of noise. The technique is completely free from any category as well as harnesses such web-contents that are weakly supervised.

A unique work carried out by Lister [42] has considered a use-case of improving knowledge transfer using machine learning approach. The idea of this model is to make use of all the essential geo-spatial information associated with educational system and use them for constructing content, searching relevant contents, and exploring essential knowledge contents. This process exponentially facilitates for SEO implementation over educational system. Adoption of SEO towards education system is also investigated by Peralta et al. [43] where a problem associated with tedious search process by teacher in finding appropriate content is addressed. The study has used a probability-based computational framework followed by resource classification in the form of clusters to make the search easier. Studies towards educational system further continues in the work of Rahman and Abdullah [44] which deals with more about customization of recommendation system.

Credibility is another essential attribute to be considered during SEO operation in order to assess the source of information. Such motive is seen to be implemented in work of Mahmood et al. [45] where reputation computation is carried out by eliminating the negative referrals. The study has used feedback-based Bayesian network in order to compute the level of expertise. Further, the work of Massaro et al. [46] have used neural network along with LSTM in order to assess the influence of web-content over an experience of user. Social network plays a dominant role in its interactive web-content where SEO plays a significant challenge to promote information on such platform in presence of complicated connected nodes in social network. Such problem is addressed in Abu-Salih et al. [47] where it targets to find the social influencer on the basis of domain considering both machine learning and semantic analysis. Further study towards social network is also seen in work of Tey et al. [48] and Xu et al. [49] where a recommendation system is built. The work carried out by Serrano [50] has investigated the impact of deep learning for computing the learning relevance towards searching voluminous web-content. It is to be noted that a structured corpora is required for building effective SEO as noted in work of Tahir et al. [51]. The work carried out by Yuan et al. [52] have used a supervised learning approach for feature normalization in order to improvise the interaction process of web contents. Further work is also carried out by Zhou et al. [53] towards user preference and recommendation of video tags is carried out by Zhou et al. [54]. Table II highlights the summary of the work being discussed in this section.

TABLE II. SUMMARY OF SEO USING ML TECHNIQUES

Authors	Problems	Techniques	Advantage	Limitation
Boppana and Sandhya [35]	Error minimization during recommendation	RNN, clustering	Achieves 99.6% of accuracy	Domain specific implementation
Burgess et al. [36]	Malicious redirection	LSTM	Achieves 98.78% of accuracy	Induces to spatial complexity
Liu and Fu et al. [37]	Identification of insecure web-contents	Unsupervised feature learning	Achieves more than 95% of precision	Induces complexity associated with feature matching during validation
Soliman et al. [38]	Effective search of web-contents	Random forest	Achieves 92% of accuracy	Doesn't address prediction performance of retrieval of data
Ismail et al. [39]	Unstructured web-contents	Fuzzy Logic	Achieves 94% of accuracy	Higher dependency towards ruleset
Hisano et al. [40]	Prediction (use-case based)	Building network with heterogeneous data, weight learning	Improved accuracy	Case specific prediction of web-contents
Kim et al. [41]	Identification for highlights of multimedia file system	Deep learning using ranking	Category independent	Iterative process, not applicable for active SEO tool
Lister [42]	Adoption of knowledge contents	Pedagogy-based learning	Helpful for knowledge delivery system	The model lacks adoption of constraints
Peralta et al. [43]	Complex search process of educational content	Recommendation system using Probability, annotation of learning resources	Better performance for hybrid recommendation	No benchmarking computationally
Rahman and Abdullah [44]	Customization of educational contents	Profile-based learning system, decision tree	Effective learning outcomes on real-test	No benchmarking
Mahmood et al. [45]	Credibility analysis	Bayesian network (Feedback)	Good convergence performance	Applicable for smaller network of web.
Massaro et al. [46]	Intelligent score allocation of webpage	LSTM, Neural network	Simplified modelling	Restricted to smaller number of webpages
Abu-Salih et al. [47]	Extracting contextual contents of social network	Sentiment analysis, machine learning, retrieval of influencer, graphical approach	Capable of processing larger data	Domain-dependent implementation
Tey et al. [48]	Recommendation issue in social network	Personalized recommender	Simplified structural model	Not applicable for complex network
Xu et al. [49]	Personalized search and recommendation	Ontological similarity	Disambiguation in recommendation design	Model dependent on human intervention towards input feature
Serrano [50]	Investigational study towards neural network	Review study towards ranking and relevance of learning models	Random neural network to have higher scope	Doesn't specifically considered internal processing of SEO
Tahir et al. [51]	Reliable corpora building	Generation of corpora	Mean yield of crawling improves significantly	Specific to language
Yuan et al. [52]	Optimizing interaction of web contents	Supervised learning	Energy reduction	Assessed on one type of client application
Zhou [53]	Evaluation of ranking performance	Gain attribute learning	Minimize dependencies on labelling	Highly iterative scheme
Zhou [54]	Recommendation (video tag)	Deep learning	Scalable model	For smaller data

## V. EXISTING SEO TOOLS

At present, there are various commercially available SEO tools which are meant to productively use time and effort towards performing data analysis and research. Some of the existing SEO tools commercially used are as follows:

### A. Commercial SEO Tools

- *Ubersuggest*: This is a free SEO tool that is meant to determine the best suited keywords followed by concluding the intention behind it. It does so by exhibiting both the long and short phrases of top ranked webpage. An exclusive report is generated on the basis of trend analysis, degree of competition, and quantity of keywords [55].
- *Moz Pro*: This SEO tool is considered as one of the best products by experts owing to its up-to-date services even compared to Google services with search algorithm. Various beneficial response are facilitated to the user via its recommendation system. Apart from this, it also offers recommendation of various keywords that contribute towards increasing page ranking. Various web-metrics are retrieved from client application in order to assess its performance via this SEO tools [56].
- *KWFinder*: The prime motive of this SEO tool is to assist in evaluating all the keywords with long trail that has minimal competitive level. It can perform evaluation of ranking as well as enhancement of specific key metric in order to upgrade popularity of webpage [57].
- *SEMRush*: This is one of the most frequently used digital marketing tool which facilitates the user to verify the ranking of their webpage. It also performs feasibility analysis for new ranking as well as analysis among different domain. Therefore, it offers significant privilege to assess their services with that of competitors on the basis of analytical report [58].
- *Google Search Console*: This tool is freely available for all users facilitated by Google. This tool can be used for indexing the sitemap of the webpage by adding their code or via using Google Analytics. This SEO tool also let the user control about the indexing policies as well as it also controls the representation structure of the website. Apart from this, the complete visualization and usage aspect of the user can be controlled by this SEO tools [59].
- *Ahrefs*: This SEO tool is mainly used for online crawling of the websites. The core purpose of its usage resides in finding out the backlinks used by the competitor. Further, it is also used for exploring the contents with highest links as well as it can also repair the broken links to find out popular web-contents [60].
- *Serpstat*: This tool is used as a hacking platform for achieving goals of content marketing and SEO. It carries out all the task that is required for managing team to analyze the competitors. It also has an enriched

availability of competitor analyzed data as well as all the aggregated keywords [61].

- There are also various other commercially available SEO tools e.g., *Screaming Frog* [62], *Keywords Everywhere* [63], *Fat Rank* [64], *Siteliner* [65], *SEOquake* [66], *Google Trends* [67], *Majestic* [68], *Woorank* [69], *SpyFu* [70], etc. Further, information about the beneficial and limiting attributes of all the discussed commercially used SEO tools are as follows:

### B. Beneficial Attributes Existing SEO Tools

The first advantage of majority of these SEO tools are that they are free of cost. The paid tools are based on usage patterns. Majority of them are reported to use local SEO tools in order to optimize the localized traffic. They are also mobile friendly as well as customer friendly while the recommendation services are based on experts.

### C. Limiting Attributes Existing SEO Tools

A robust usage of SEO will yield a page with higher rank and this will also attract the attention of competitors. Hence, this is a continuous effort to be at top of rank, which is extremely challenging. There are fair feasibility of SEO to change which often causes uncertainty of consistency of ranks in upcoming times. The process of generation of response in SEO is quite a slower process. Even after frequent webpage updating, there is no assurance of timely results within a tentative duration of time.

## VI. EXISTING RESEARCH TREND

At present, there are different categories of studies being undertaken for improving the performance of SEO. Table III highlights the research trends of using different standard approaches in SEO.

TABLE III. SUMMARY RESEARCH TRENDS ON SEP (2017-2022)

Items	Conference	Journal	Early Access Article	Books	Magazine
Total manuscript	298	70	12	3	3
NLP-based approach	2	10	0	0	0
ML-based approach	29	12	1	0	0
Miscellaneous	267	48	11	3	3

From Table III, it can be seen that there are very a smaller number of journal publications associated with both standard NLP and ML based approach in SEO as compared to miscellaneous approaches, which are normally application specific. The trend of minimal journal publication eventually means that both NLP and ML approach has just very a smaller number of research implementation in IEEE Xplore digital library. Similar trend of publication towards NLP and ML is also observed for other reputed publication of Elsevier, Springer, Wiley, etc. This concludes that there should be more attempts towards wholesome utilization of NLP and ML

approach for addressing the open-end research problems as illustrated in next section.

## VII. RESEARCH GAP

After reviewing the existing approaches towards addressing the challenges in SEO, following research gap has been identified.

### A. *More Focus on Local Problems*

A closer look into the existing approaches towards SEO shows that there are different variants of techniques in order to address specific set of problems or to cater up certain application demands. However, there is no existing framework, which can develop a solution towards addressing combined local problems over webpages e.g., duplicated contents, difference in performance in different computing device (e.g., PC, tablet, Smartphone), poor link building, inaccurate navigation system, not search friendly, inaccurate redirection, cluttered URLs, loading of page to be consuming high time, ignoring local search or not considering markup data. Although, all the above-mentioned problems have been individually found to be investigated, but they have not been combinedly addressed. Solving some of the local problems and ignoring the remaining of problems will eventually lead to impractical solution towards improving SEO.

### B. *Few Emphases Towards Content Generation*

One of the targets of the SEO approach is to generate a precise content in order to meet the business objectives by reaching to maximum targeted customer. However, this is highly computationally challenging task. Existing approaches has evolved up with various techniques to ensure content quality, meta-data generation, and accuracy in its predictive approach. Such problems are mainly found to be solved using different variants of artificial intelligence and ML approaches. However, all such ML techniques suffer from serious drawbacks either of computational complexities or towards dependencies toward massive trained data. Existing ML approaches are also highly iterative and is mainly meant for passive mode of predictive operation. Therefore, they are less likely to be used for practical world application of SEO.

### C. *Few Studies Towards Smart Content Management*

There is no doubt that ranking plays a significant role in SEO building process. However, such forms of ranking mechanism suffer from lower scale of adoption of objective function. Moreover, usage of existing deep learning scheme makes the process so much complicated and resource dependent that there is less scope of performing updating procedure. Without proper updating procedure, it is impossible to revise the solution being built for addressing local problems in SEO. At the same time, implementation of existing frameworks using NLP or ML will require serious re-engineering process, which is definitely not a cost-effective deployment scheme.

Hence, all the above-mentioned research gap are required to be bridged, without which a better form of SEO tool is impractical to be designed.

## VIII. DISCUSSION AND RESEARCH IMPLICATIONS

In this survey work, the study explored the use of NLP and ML techniques in SEO. Through the literature review, it has been found that NLP techniques are particularly useful in improving the readability and quality of web content, while ML techniques are effective in analyzing various factors that influence search rankings. However, a combination of both techniques is often most effective, and there is a growing body of research on the integration of NLP and ML in SEO. This section delves deeper into the specific implications of these findings. The entire section includes discussing the practical implications of these findings for SEO practitioners and web content creators. Additionally, this section addresses challenges in the current research on NLP and ML for SEO, and suggests potential avenues for future research to address these issues.

### A. *Findings and Discussion*

One of the most significant challenges in SEO is predicting and analyzing search engine algorithms. Based on the above-mentioned discussion it has been explored that, both NLP and ML techniques have been increasingly used in SEO to help search engines better understand the intent and meaning of web content, and to improve search rankings. One of the most common applications is the use of NLP to better understand search queries and match them with relevant content. It can be adopted to identify the underlying meaning and intent of search queries, and then match them with the most relevant content on the web. Another way that NLP techniques can be used in SEO is to improve the readability and quality of web content. Researchers have developed tools that use NLP to analyze the readability, grammar, and spelling of web content, and provide suggestions for improvement. On the other hand, ML techniques have been also be used to improve SEO in a number of ways. One of the most common applications is the use of ML to predict search rankings. Researchers have developed algorithms that use ML to analyze various factors that influence search rankings, such as keyword density, backlinks, and user engagement, and then make predictions on which websites are most likely to rank highly.

Irrespective of various number of research-based models being evolved, there are still an open-end problems associated with the performance of SEO. From commercial application viewpoint, existing studies don't promote towards potential links exploration while developing the model which will present the client webpage towards maximized rankings of search engine. The existing models do offer some solution to promote the popular content based on domain specific frameworks; however, there is lack of consistency towards the link building process. Irrespective of various study implementation using NLP, existing research work also doesn't seem to consider much of content management programs along with considering complexities of data within it. One such issue is presence of iterative tags of title, which still existing NLP is not able to address properly. The content management using NLP is required to be consistently updated, without which dynamic crawling could lead to ineffective convergence of search operation of web contents. Although, existing contribution of ML are quite notable; but they are also scattered as well as highly specific to use-cases. Hence, adoption of such models will be quite expensive and will require time-to-time

update and amendment based on business structure. At present, there is no generalized architecture or framework to address global problems all together. There is an increased proliferation in using different variants of ML approach towards optimizing various essential operation in building an effective SEO. However, existing ML approaches are mainly iterative, demands voluminous set of data, and doesn't have much consideration of multi-objective function along with adoption of practical constraints. This further reduces the scope of predictive approach and hence, existing SEO has not yet harnessed the full capabilities of ML approaches in order to gain a better result.

### B. Remarks and Implications

It's difficult to determine which method is better, as both NLP and ML have their own strengths and weaknesses, and the choice of method will depend on the specific application and context.

NLP techniques are particularly useful in understanding the natural language used in search queries and web content. They can help search engines better understand the intent and meaning behind search queries, and can also improve the readability and quality of web content. However, NLP techniques may not be as effective in analyzing more quantitative factors, such as keyword density and backlinks, which are important for search rankings. On the other hand, ML techniques are particularly useful in analyzing large amounts of data and identifying patterns that are difficult for humans to detect. They can be used to analyze various factors that influence search rankings, such as keyword density, backlinks, and user engagement, and can make predictions on which websites are most likely to rank highly. However, ML techniques may not be as effective in analyzing the natural language used in search queries and web content.

While NLP and ML techniques are often used separately in SEO, there is also a growing body of research on the integration of these techniques. In many cases, a combination of both NLP and ML techniques may be most effective. For example, using NLP to better understand search queries and match them with relevant content, and using ML to predict search rankings based on a range of factors. Additionally, the effectiveness of either technique will depend on the quality of the data used and the specific algorithms and models used. Another area of research is the use of NLP and ML to identify and address black hat SEO techniques, such as keyword stuffing and link farming. An algorithm can be developed using NLP and ML to detect web content that has been artificially optimized for search engines and prevent websites from using them to manipulate search rankings.

Although, the use of NLP and ML in SEO offers numerous benefits, but it also has limitations, including the accuracy of the algorithms used and the cost of implementing technology. As ML and NLP technology continues to advance, it is likely that it will become increasingly essential in optimizing website ranking and visibility.

## IX. CONCLUSION

This paper has investigated towards the performance improvement approaches from research viewpoint towards

developing a strong ecosystem of a holistic marketing. In this perspective, there are evolution of massive number of searches by the customers and digital marketers annually with an intention of fulfilling certain commercial targets. The prime outcome is to end up their search towards more relevant conclusive services or products. For this purpose, the webpage is required to be optimized for maximized ranking and higher visibility. Based on above learning outcomes, potential research gap is explored and the future work will be carried out towards addressing all the pitfalls of existing system as well as adopt all the beneficial points of the existing literatures. The first research gap is possible to be solved by developing a unique architecture integrating both NLP and ML approach, which will be capable to address majority of the local problems in building SEO using predictive page ranking approach. The second research gap can be solved by further improving the similar architecture and add novel functionalities towards efficient content generation process in SEO. A new variant of deep learning approach can be used with feedback connection over a tree-based network system. This will offer a capability to processes complete sequence of data available in web page. Focus will be also towards achieving better predictive generated data with lesser epoch values for confirming lower computational complexities. The third research gap can be addressed by further improving the same model using improved version of machine learning algorithm. In order to meet an optimization objective, a multi-objective function can be designed using three parameters i.e., state, reward, and actions in order to get more updated contents.

## REFERENCES

- [1] N. Papagiannis, *Effective SEO and Content Marketing The Ultimate Guide for Maximizing Free Web Traffic*, Wiley, ISBN: 9781119628859, 1119628857, 2020
- [2] A. Veglis, D. Giomelakis, *Search Engine Optimization*, MDPI AG, ISBN: 9783039368181, 3039368184, 2021
- [3] T. Kelsey, *Introduction to Search Engine Optimization-A Guide for Absolute Beginners*, Apress, ISBN: 9781484228517, 1484228510, 2017
- [4] L. Welz, *SEO For Beginners-Explained SEO In Simple Language, Beginner To Advanced: Marketing Strategies Book*, Independently Published, ISBN: 9798714191374, 2021
- [5] J. Knight, *SEO For Beginners 2020-Learn and Develop a Strategy for Search Engine Optimization and Grow Your Business With Google*, Amazon Digital Services LLC - KDP Print US, ISBN: 9781670861061, 1670861066, 2019
- [6] A. V. Patil and V. Madhukar Patil, "Search Engine Optimization Technique Importance," 2018 IEEE Global Conference on Wireless Computing and Networking (GCWCN), 2018, pp. 151-154, doi: 10.1109/GCWCN.2018.8668581.
- [7] V. M. Patil and A. V. Patil, "SEO: On-Page + Off-Page Analysis," 2018 International Conference on Information , Communication, Engineering and Technology (ICICET), 2018, pp. 1-3, doi: 10.1109/ICICET.2018.8533836.
- [8] A. Husayni, *The Google SEO Handbook-How to Analyze and Optimize Your Site's Search Footprint Like a Pro, Millionairium*, ISBN: 9780990782001, 099078200X, 2019
- [9] V. Duong, *Baidu SEO-Challenges and Intricacies of Marketing in China*, Wiley, ISBN: 9781119368724, 1119368723, 2017
- [10] K. Sandhu, *Emerging Challenges, Solutions, and Best Practices for Digital Enterprise Transformation*, Business Science Reference, ISBN: 9781799885894, 1799885895, 2021
- [11] Van Looy A. "Search Engine Optimization". In: *Social Media Management*. Springer Texts in Business and Economics. Springer, Cham. (2016), [https://doi.org/10.1007/978-3-319-21990-5\\_6](https://doi.org/10.1007/978-3-319-21990-5_6)

- [12] V. N. Gudivada, D. Rao and J. Paris, "Understanding Search-Engine Optimization," in *Computer*, vol. 48, no. 10, pp. 43-52, Oct. 2015, doi: 10.1109/MC.2015.297
- [13] Z. Hui, Q. Shigang, L. Jinhua and C. Jianli, "Study on Website Search Engine Optimization," 2012 International Conference on Computer Science and Service System, 2012, pp. 930-933, doi: 10.1109/CSSS.2012.236
- [14] Hussien, A. S. "Factors affect search engine optimization." *International Journal of Computer Science and Network Security* 14, no. 9 (2014): 28-33.
- [15] Persynska, K.: 8 risky black hat SEO techniques used today. Positionally Blog (2015)
- [16] Agrawal, S., Somani, A., Chhabra, V.: Discernment of search engine spamming and counter measure for it, India, 8 August 2016
- [17] Killoran, John B. "How to use search engine optimization techniques to increase website visibility." *IEEE Transactions on professional communication* 56, no. 1 (2013): 50-66.
- [18] Hajeer, Safaa I., Rasha M. Ismail, Nagwa L. Badr, and Mohamed Fahmy Tolba. "A new stemming algorithm for efficient information retrieval systems and web search engines." In *Multimedia Forensics and Security*, pp. 117-135. Springer, Cham, 2017.
- [19] Tsuei, Hung-Jia, Wei-Ho Tsai, Fu-Te Pan, and Gwo-Hsiung Tzeng. "Improving search engine optimization (SEO) by using hybrid modified MCDM models." *Artificial Intelligence Review* 53, no. 1 (2020): 1-16.
- [20] Luh, Cheng-Jye, Sheng-An Yang, and Ting-Li Dean Huang. "Estimating Google's search engine ranking function from a search engine optimization perspective." *Online Information Review* (2016).
- [21] Jenkins, Porter, Jennifer Zhao, Heath Vinicombe, Anant Subramanian, Arun Prasad, Atillia Dobi, Eileen Li, and Yunsong Guo. "Natural language annotations for search engine optimization." In *Proceedings of The Web Conference 2020*, pp. 2856-2862. 2020.
- [22] Sharma, Sunny, Sunita Mahajan, and Vijay Rana. "A semantic framework for ecommerce search engine optimization." *International Journal of Information Technology* 11, no. 1 (2019): 31-36.
- [23] Zhang, Lu, Tiantian Wang, Yiran Liu, and Qingling Duan. "A semi-structured information semantic annotation method for Web pages." *Neural Computing and Applications* 32, no. 11 (2020): 6491-6501.
- [24] F. Horasan, "Keyword extraction for search engine optimization using latent semantic analysis," *J. Polytech.*, 24, no. 2: 473-479 2020.
- [25] E. Uzun, "A Novel Web Scraping Approach Using the Additional Information Obtained From Web Pages," in *IEEE Access*, vol. 8, pp. 61726-61740, 2020, doi: 10.1109/ACCESS.2020.2984503.
- [26] D. P. Barrett, A. Barbu, N. Siddharth and J. M. Siskind, "Saying What You're Looking For: Linguistics Meets Video Search," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 10, pp. 2069-2081, 1 Oct. 2016, doi: 10.1109/TPAMI.2015.2505297.
- [27] D. Dominguez-Sal, J. Aguilar-Saborit, M. Surdeanu and J. L. Larribapey, "Using Evolutionary Summary Counters for Efficient Cooperative Caching in Search Engines," in *IEEE Transactions on Parallel and Distributed Systems*, vol. 23, no. 4, pp. 776-784, April 2012, doi: 10.1109/TPDS.2011.162.
- [28] A. Ghanbarpour and H. Naderi, "An Attribute-Specific Ranking Method Based on Language Models for Keyword Search over Graphs," in *IEEE Transactions on Knowledge and Data Engineering*, vol. 32, no. 1, pp. 12-25, 1 Jan. 2020, doi: 10.1109/TKDE.2018.2879863.
- [29] S. Soltani, S. A. H. Seno and R. Budiarto, "Developing Software Signature Search Engines Using Paragraph Vector Model: A Triage Approach for Digital Forensics," in *IEEE Access*, vol. 9, pp. 55814-55832, 2021, doi: 10.1109/ACCESS.2021.3071795.
- [30] X. Dai, H. Dai, G. Yang, X. Yi and H. Huang, "An Efficient and Dynamic Semantic-Aware Multikeyword Ranked Search Scheme Over Encrypted Cloud Data," in *IEEE Access*, vol. 7, pp. 142855-142865, 2019, doi: 10.1109/ACCESS.2019.2944476.
- [31] L. -C. Chen, "A Study of Optimizing Search Engine Results Through User Interaction," in *IEEE Access*, vol. 8, pp. 79024-79045, 2020, doi: 10.1109/ACCESS.2020.2990972.
- [32] F. Zhang, H. Niu, I. Keivanloo and Y. Zou, "Expanding Queries for Code Search Using Semantically Related API Class-names," in *IEEE Transactions on Software Engineering*, vol. 44, no. 11, pp. 1070-1082, 1 Nov. 2018, doi: 10.1109/TSE.2017.2750682.
- [33] E. A. Calvillo, R. Mendoza, J. Munoz, J. C. Martinez, M. Vargas and L. C. Rodriguez, "Automatic algorithm to classify and locate research papers using natural language," in *IEEE Latin America Transactions*, vol. 14, no. 3, pp. 1367-1371, March 2016, doi: 10.1109/TLA.2016.7459622.
- [34] E. Hamzei and F. Hakimpour, "Entity recognition and disambiguation for natural-language spatial search queries," 2017 3th International Conference on Web Research (ICWR), 2017, pp. 32-37, doi: 10.1109/ICWR.2017.7959301.
- [35] V. Boppana & P. Sandhya, "Web crawling based context aware recommender system using optimized deep recurrent neural network", *SpringerOpen-Journal of Big Data*, Article No. 144, 2021
- [36] J. Burgess, P. O'Kane, S. Sezer and D. Carlin, "LSTM RNN: detecting exploit kits using redirection chain sequences", *SpringerOpen-Cybersecurity*, Article No. 25, 2021
- [37] X. Liu and J. Fu, "SPWalk: Similar Property Oriented Feature Learning for Phishing Detection," in *IEEE Access*, vol. 8, pp. 87031-87045, 2020, doi: 10.1109/ACCESS.2020.2992381.
- [38] H. Soliman, "Random Forest Based Searching Approach for RDF," in *IEEE Access*, vol. 8, pp. 50367-50376, 2020, doi: 10.1109/ACCESS.2020.2980155.
- [39] H. M. Ismail, B. Belkhouche and S. Harous, "Framework for Personalized Content Recommendations to Support Informal Learning in Massively Diverse Information Wikis," in *IEEE Access*, vol. 7, pp. 172752-172773, 2019, doi: 10.1109/ACCESS.2019.2956284.
- [40] R. Hisano, D. Sornette, and T. Mizuno, "Prediction of ESG compliance using a heterogeneous information network", *SpringerOpen-Journal of Big Data*, Article No. 22, 2020
- [41] H. Kim, T. Mei, H. Byun and T. Yao, "Exploiting Web Images for Video Highlight Detection With Triplet Deep Ranking," in *IEEE Transactions on Multimedia*, vol. 20, no. 9, pp. 2415-2426, Sept. 2018, doi: 10.1109/TMM.2018.2806224.
- [42] P. J. Lister, "A smarter knowledge commons for smart learning", *SpringerOpen-Smart Learning Environment*, Article No 8, 2018
- [43] M. Peralta, R. Alarcon, K. Pichara, T. Mery, F. Cano and J. Bozo, "Understanding Learning Resources Metadata for Primary and Secondary Education," in *IEEE Transactions on Learning Technologies*, vol. 11, no. 4, pp. 456-467, 1 Oct.-Dec. 2018, doi: 10.1109/TLT.2017.2766222.
- [44] M. M. Rahman and N. A. Abdullah, "A Personalized Group-Based Recommendation Approach for Web Search in E-Learning," in *IEEE Access*, vol. 6, pp. 34166-34178, 2018, doi: 10.1109/ACCESS.2018.2850376.
- [45] S. Mahmood, A. Ghani, A. Daud and S. Shamshirband, "Reputation-Based Approach Toward Web Content Credibility Analysis," in *IEEE Access*, vol. 7, pp. 139957-139969, 2019, doi: 10.1109/ACCESS.2019.2943747.
- [46] A. Massaro, D. Giannone, V. Birardi and A. M. Galiano, "An Innovative Approach for the Evaluation of the Web Page Impact Combining User Experience and Neural Network Score", *MDPI Journal, Future Internet*, vol.12, Iss.145, 2021.https://doi.org/10.3390/fi13060145
- [47] B. Abu-Salih, K. Y. Chan, O. Al-Kadi, "Time-aware domain-based social influence prediction", *SpringerOpen-Journal of Big Data*, Article No.10, 2020
- [48] F. J. Tey, T-Y Wu, C-L Lin, and J-L Chen, "Accuracy improvements for cold-start recommendation problem using indirect relations in social networks", *SpringerOpen-Journal of Big Data*, vol.8, Iss.98, 2021
- [49] Z. Xu, O. Tifrea-Marcuska, T. Lukaszewicz, M. V. Martinez, G. I. Simari and C. Chen, "Lightweight Tag-Aware Personalized Recommendation on the Social Web Using Ontological Similarity," in *IEEE Access*, vol. 6, pp. 35590-35610, 2018, doi: 10.1109/ACCESS.2018.2850762
- [50] W. Serrano, "Neural Networks in Big Data and Web Search", *MDPI, data*, vol.4, Iss.7, 2019.doi:10.3390/data4010007

- [51] B. Tahir and M. A. Mehmood, "Corpulyzer: A Novel Framework for Building Low Resource Language Corpora," in *IEEE Access*, vol. 9, pp. 8546-8563, 2021, doi: 10.1109/ACCESS.2021.3049793.
- [52] L. Yuan, J. Ren, L. Gao, Z. Tang and Z. Wang, "Using Machine Learning to Optimize Web Interactions on Heterogeneous Mobile Systems," in *IEEE Access*, vol. 7, pp. 139394-139408, 2019, doi: 10.1109/ACCESS.2019.2936620.
- [53] K. Zhou, H. Zha, Y. Chang and G. -R. Xue, "Learning the Gain Values and Discount Factors of Discounted Cumulative Gains," in *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 2, pp. 391-404, Feb. 2014, doi: 10.1109/TKDE.2012.252
- [54] R. Zhou, D. Xia, J. Wan and S. Zhang, "An Intelligent Video Tag Recommendation Method for Improving Video Popularity in Mobile Computing Environment," in *IEEE Access*, vol. 8, pp. 6954-6967, 2020, doi: 10.1109/ACCESS.2019.2961392.
- [55] Patel, Neil. "Ubersuggest: Free Keyword Research Tool." Available: <https://neilpatel.com/ubersuggest>. [Accessed: 21-Nov-2022]
- [56] Moz. (n.d.). Moz Pro. Retrieved from <https://moz.com/products/pro>. [Accessed: 22-Nov-2022]
- [57] KWFinder. (n.d.). Keyword Research and Analysis Tool. Retrieved from <https://kwfinder.com/>. [Accessed: 22-Nov-2022]
- [58] Semrush - online marketing can be easy," Semrush. [Online]. Available: <https://www.semrush.com/>. [Accessed: 22-Nov-2022].
- [59] Google Search Central (formerly Webmasters)," Google Developers. [Online]. Available: <https://developers.google.com/search>. [Accessed: 22-Nov-2022]
- [60] "Ahrefs - SEO tools & resources to grow your search traffic," Ahrefs.com. [Online]. Available: <https://ahrefs.com/>. [Accessed: 22-Nov-2022]
- [61] "Serpstat — growth hacking tool for SEO, PPC and content marketing," Serpstat.com. [Online]. Available: <https://serpstat.com/>. [Accessed: 22-Nov-2022].
- [62] D. T. C. Stewart, Ed., *Screaming Frog Seo Spider*. Dicho, 2012.
- [63] "Browser add-on to see Google search volume everywhere," Keywordseverywhere.com. [Online]. Available: <https://keywordseverywhere.com>. [Accessed: 22-Nov-2022]
- [64] "FatRank - digital nomad," FatRank, 31-Jul-2017. [Online]. Available: <https://www.fatrank.com/>. [Accessed: 22-Nov-2022].
- [65] "Siteliner - Find Duplicate Content on your site," Siteliner.com. [Online]. Available: <https://www.siteliner.com/>. [Accessed: 22-Nov-2022].
- [66] A Powerful SEO Toolbox for your Browser," Seoquake.com. [Online]. Available: <https://www.seoquake.com/index.html>. [Accessed: 22-Nov-2022].
- [67] "Google trends," Google Trends. [Online]. Available: <https://trends.google.com/trends/?geo=IN>. [Accessed: 22-Nov-2022]
- [68] Majestic maps and categorizes the web," Majestic.com. [Online]. Available: <https://majestic.com/>. [Accessed: 22-Nov-2022].
- [69] "Website optimization and digital agency sales tools," Woorank.com. [Online]. Available: <https://www.woorank.com/>. [Accessed: 22-Nov-2022].
- [70] K. JFounder/CEO, "SpyFu - competitor keyword research tools for Google ads PPC & SEO," Spyfu.com. [Online]. Available: <https://www.spyfu.com/>. [Accessed: 22-Nov-2022].

# Adaptive Rectified Linear Unit (AReLU) for Classification Problems to Solve Dying Problem in Deep Learning

Ibrahim A. Atoum

Department of Computer Science and Information Systems  
College of Applied Sciences, Al Maarefa University  
Riyadh, Saudi Arabia

**Abstract**—A convolutional neural network (CNN) is a subset of machine learning as well as one of the different types of artificial neural networks that are used for different applications and data types. Activation functions (AFs) are used in this type of network to determine whether or not its neurons are activated. One non-linear AF named as Rectified Linear Units (ReLU) which involves a simple mathematical operations and it gives better performance. It avoids rectifying vanishing gradient problem that inherents older AFs like tanh and sigmoid. Additionally, it has less computational cost. Despite these advantages, it suffers from a problem called Dying problem. Several modifications have been appeared to address this problem, for example; Leaky ReLU (LReLU). The main concept of our algorithm is to improve the current LReLU activation functions in mitigating the dying problem on deep learning by using the readjustment of values (changing and decreasing value) of the loss function or cost function while number of epochs are increased. The model was trained on the MNIST dataset with 20 epochs and achieved lowest misclassification rate by 1.2%. While optimizing our proposed methods, we received comparatively better results in terms of simplicity, low computational cost, and with no hyperparameters.

**Keywords**—Rectified Linear Unit (ReLU); Convolutional Neural Network; activation function; deep learning; MNIST dataset; machine learning

## I. INTRODUCTION

The concept of Artificial Intelligence (AI) revolves around creating intelligent machines that are able to simulate human thinking while Machine Learning (ML) is a branch of this concept that allows these intelligent machines to learn the hidden patterns from the input data [1]. Neural networks (NNs) as a subset of ML simulate the human brain using a set of algorithms. These networks consist of input, hidden, and output layers. These layers consist of neurons that mimic the structure of a biological neuron, where each neuron has inputs that are processed to give outputs, which in turn will be input to another neuron. When neural networks consist of more than three layers then they can be called Deep Learning Networks (DLNs) [2].

AFs play a critical role in DLNs to extract the results from the input values and thus determine whether the underline neuron is activated or not [3]. DLNs can be considered as just a linear regression without AFs, so appropriate AFs must be used to model a nonlinear DLNs. AFs classified as binary step, linear activation and nonlinear activation functions.

Binary step function is a basic threshold classifier where some threshold value is decided to choose which output neurons should be activated or deactivated. Linear activation function is a simple straight line activation function that converts linear input signals into non-linear output signals. Nonlinear AFs are what make it easier for the DLNs model to adapt to a variety of data and to distinguish between outcomes; examples are: ReLU, Leaky ReLU, Sigmoid, Tanh and Softmax [4]. Some of these are suited to be used in hidden layers and others in output layers.

There are two terms used in training the model, the first is the term feedforward, which is used in NNs to refer to the transition with specified weights from input to output, while the term backpropagation, as the name suggests, moves from output to input with readjustment of weights depending on loss values and then propagation processes straight ahead. This approach allows the use of gradient methods, such as gradient descent or stochastic gradient descent, to train multi-layer networks and update weights to reduce loss [5] [6]. ReLU as a nonlinear AF has gained a lot of interest in research due to its simplicity, low computation cost and it avoids the vanishing gradient problem that inherent to the earliest AFs like tanh and sigmoid [7]. Despite all the previous advantages of this function, it has a problem called the Dying ReLU problem, which indicates that the neuron becomes inactive and outputs zero only for any input. This problem has been attributed to a high learning rate and a high negative bias [8].

ReLU was initially introduced by [9]; the researchers designed an electronic circuit to simulate a hybrid slug in which the latent cortex combines the digital selection of an active cluster of neurons with an analog response, and this behavior is achieved by dynamically changing the positive feedback inherent in recurrent cortical connections, this behavior, according to the researchers, created computational capabilities creates the process of stimulus selection, conferring the ability to modify and generate a spatio-temporal pattern in this cortex.

ReLU was later used in object recognition by [10], and researchers summarized the three stages used to extract object features such as filter bank, nonlinear transformation and a kind of feature pooling layer emphasizing that most systems use one or two of these stages, assuming that the use of two stages gives more accurate results. The study demonstrated the

accuracy of this hypothesis by using nonlinear layers and pooling layers on different object data sets through either supervised optimization or unsupervised pre-training.

ReLU was also popularized by [11] in the context of Restricted Boltzmann Machines. The study demonstrated how to create a more powerful type of hidden units for Restricted Boltzmann Machines (RBM) in object recognition and face comparison by combining weights and biases for an infinite set of binary units with approximating these stepped sigmoid units with noisy corrected linear units.

Leaky ReLU [12] has added a slight slope in the negative range; this modification on ReLU ends the presence of dead neurons in the negative region by using a hyperparameter. Thereafter many leaky ReLU variants have been appeared like Parametric Rectified Linear Unit (PReLU) [13] which introduces a new learnable parameter as a slope for the negative part and Exponential linear unit (ELU) has used an exponential function to transition from the positive to small negative values [14].

The value of the loss function is related to the results of the model. If the value of the loss function is low, this means that the model will give good results [16]. Loss functions are divided into two types, classification and regression. Classification functions also divided into binary entropy loss/log loss and hinge loss. During the execution of AReLU the first function was used [15]. AReLU is applied on MNIST dataset that contains 70000 images of black and white handwritten digits divided into 60000 images for training and 10000 images for Testing [17].

In this study, the decreasing value of the used loss function was exploited as an adaptive parameter to keep the network active. The study is presented into four sections: section two introduces the idea of ReLU, section three identifies the ReLU dying problem, and section four introduces the AReLU. Section five presents the results and finally section six is the conclusion.

## II. RECTIFIED LINEAR UNIT (RELU) ACTIVATION FUNCTION

Artificial neurons are mathematical model that mimic human biological neurons and they are the basic building blocks of neural networks as shown in Fig. 1.

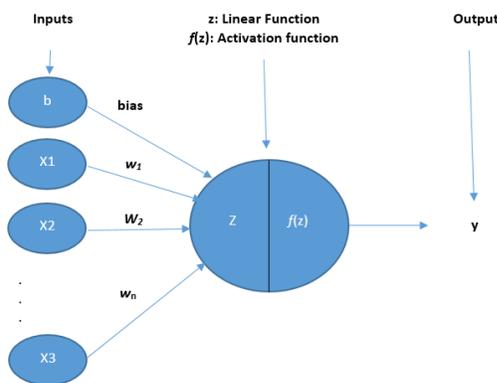


Fig. 1. Artificial neurons representation

Where  $z = ((w_1 \times x_1) + (w_2 \times x_2) + \dots \dots \dots + (w_n \times x_n)) + b$  states the linear function, accordingly the activation function  $y = f(z)$ .  $x_1$  to  $x_n$  represents the inputs,  $w_1$  to  $w_n$  illustrates the weights that connect inputs with perceptions and they measure the significance level of each input. The bias value ( $b$ ) is added to the weighted sum of the inputs to prevent the activation function from getting a zero value. This linear results in linear modeling come from the linear mapping of the input function to output in hidden layers. The role of activation function is to convert these linear outputs into non-linear outputs for further computation as in

$y = \alpha * (((w_1 \times x_1) + (w_2 \times x_2) + \dots \dots \dots + (w_n \times x_n)) + b)$ ; where  $\alpha$  is the activation function. The literature has introduced many activation functions such as Sigmoid, binary step, Tanh, ReLU, Leaky ReLU, identity and Softmax.

ReLU activation function can be described mathematically as in Eq. (1) and graphically as in Fig. 2, where  $x$  is the input to the neuron. The function  $f(x)$  equal zero for all negative input values and equal original input value for all positive input values as in Eq. (1).

$$f(x) = \max(0,x) \quad (1)$$

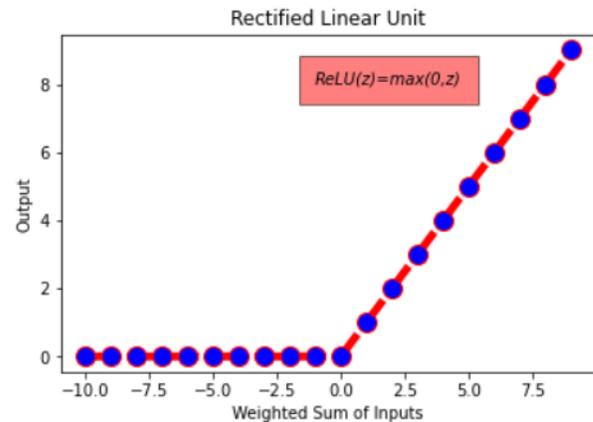


Fig. 2. ReLU representation

ReLU avoids vanishing gradient problem occurred with other activation functions by preserving the gradient [18]. This problem is formed when the gradients of deep neurons vanish or becomes zero, this means that the deep layers of the network may not learn or learn very slowly [19]. Derivative Activation function is fundamental to optimizing neural network, the ReLU ( $x$ ) can be expressed as:

$$f(x) = \max(0, x)$$

It can be simplified as follows:

$$\max(0, x) = \begin{cases} 0, & x < 0 \\ x, & x \geq 0 \end{cases}$$

The first order derivative of this function is:

$$\frac{d}{dx} f(x) = \frac{d}{dx} \max(0, x) = \frac{d}{dx} \begin{cases} 0, & x < 0 \\ x, & x \geq 0 \end{cases}$$

And can be illustrated as:

$$\frac{d}{dx}f(x) = \frac{d}{dx} \begin{cases} \frac{d}{dx}(0), x < 0 \\ \frac{d}{dx}(x), x \geq 0 \end{cases}$$

The final derivative is:

$$\frac{d}{dx}f(x) = \frac{d}{dx} \begin{cases} 0, x < 0 \\ x, x \geq 0 \end{cases}$$

### III. RELU DYING PROBLEM

Dying ReLU problem is one limitation for ReLU where its neurons output is zero as illustrated by the red outline in Fig. 3.

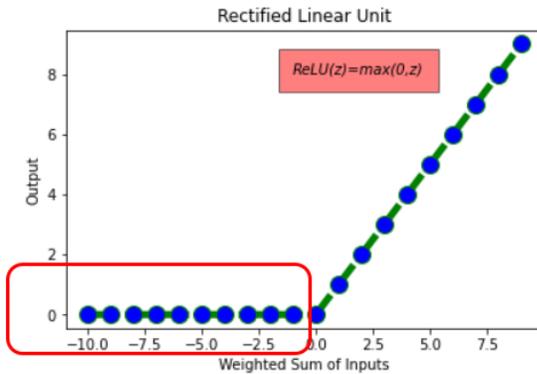


Fig. 3. Red outline where ReLU outputs 0

The normal situation for ReLU neurons is to stay active, update weights, and keep learning. Although this feature provides the power to ReLU through the sparsity of the network, it poses a problem when most of the inputs of these ReLU neurons are in the negative range, and the issue becomes more complicated when the output of most Neurons is zero, making their task so abnormal that they become inactive and unlearning. This inevitably causes gradients to fail to flow during backpropagation.

The cause of this problem is due to two main factors: High Learning Rate and a Large Negative Bias. The former one allows faster learning with the possibility of a numerical overflow, while its very small value may never converge or stumble on a suboptimal solution. So choosing an average rate that is neither too large nor too small ensures an optimal approximation of the mapping problem as represented by the training data set. The best way to discover the value of the learning rate is through trial and error, not analytically for a particular model on a particular data set. This can be illustrated by the update process in backpropagation as shown in Eq. (2).

$$newW_{ij} = oldW_{ij} - LR \left( \frac{\partial Error}{\partial oldW_{ij}} \right) (2)$$

where  $\frac{\partial Error}{\partial oldW_{ij}}$  is the derivative of error with respect to weight. We can see from Eq. (2) that giving a high value of the learning rate (LR) will cause a high value for the last part of Eq. (2)  $LR * \left( \frac{\partial Error}{\partial oldW_{ij}} \right)$ , so subtracting large number from

$oldW_{ij}$  will end up with highly negative  $newW_{ij}$ . These negative results cause negative inputs for ReLU, therefore generating the dying ReLU problem.

Biases are extra inputs that ensure neurons are activated regardless of the input. Changing the value of the weights in the neuron changes the steepness of the curve without the ability to change it to the right or left, to change the curve to the left or right the value is changed. Giving a high negative bias value makes the ReLU activation input negative. To mitigate Dying ReLU problem, several techniques have emerged, all trying to keep the network active when the input is negative or zero.

Leaky ReLU [12] demolished dead neurons in the negative part by adding a slight slop in the negative range using a hyperparameter ( $\alpha=0.1$  or more) as shown in Eq. (3) and illustrated in Fig. 4.

$$f(x) = \max(ax, x), \text{ where } a > 0 \quad (3)$$

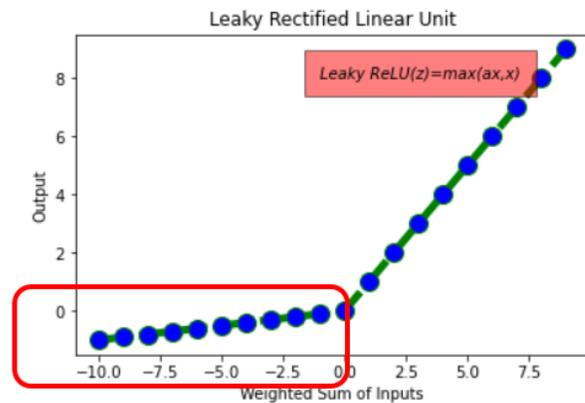


Fig. 4. Red outline where leaky ReLU outputs less than zero

Parametric Rectified Linear Unit (PReLU) [13] thereafter presented a new learnable parameter as a slope for the negative part as in Eq. (4):

$$f(x) = \max(ax, x), \text{ where } a \text{ is a Learnable Parameter} \quad (4)$$

And Exponential linear unit (ELU) used an exponential function  $\propto (e^x - 1)$  to transition from the positive to small negative values [14] as shown in Eq. (5).

$$f(x) = \begin{cases} x, x > 0 \\ \alpha (e^x - 1), x \leq 0 \end{cases} \quad (5)$$

### IV. ADAPTIVE RECTIFIER LINEAR UNIT (ARELU) ON MNIST DATASET

The study used the MNIST dataset of handwritten greyscale images, these images were size-normalized and centered in a fixed-size image available from NIST [20] as shown in Fig. 5 which shows the first 25 images of MNIST.

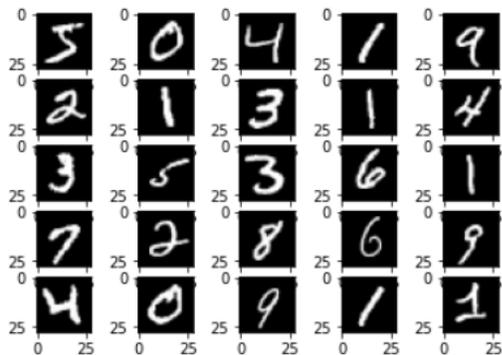


Fig. 5. Subset of MNIST dataset

This dataset composed of approximately 70,000 handwritten monochrome images of 0 to 9 (10 digits), each of which is 784 pixels in size, so that the input data is in pairs (70,000,784) and output (70,000, 10) as shown in Fig. 6.

To form the network, the AReLU activation function were used in the hidden layer and softmax in the output layer. The used loss function is categorical\_crossentropy and the optimizer is Adamax. The batch size is adopted to 128 and the number of epochs to 20.

Once the output is generated from the final neural net layer, loss function (input vs output) is calculated and backpropagation process is performed where the weights are adjusted to get the minimum loss. Neural Networks are trained using the gradient descent process. This process consists of the backward propagation step which is basically chain rule to get the change in weights in order to reduce the loss after every epoch.

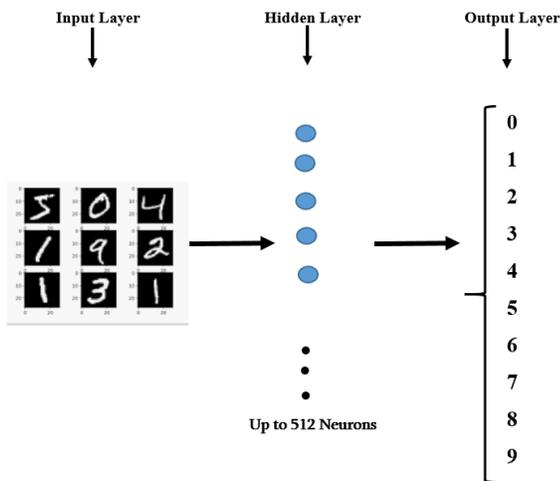


Fig. 6. MNIST neural network

The primary goal of AReLU is to mitigate Dying ReLU problem by improving the previous methods by using the adaptive Loss Function ( $l$ ) parameter instead of hyperparameter one.  $l$  is multiplied by the input value as shown in Eq. (6) to transit from the positive to small negative values.

$$f(x) = \begin{cases} x, & x > 0 \\ l * x, & x \leq 0 \end{cases} \quad (6)$$

The AReLU has implemented by using Python programming language according to the algorithm shown in Fig. 7 and more illustrated in Fig. 8. It is noticed from the equation that there is no change in the case of the positive values, but only the change in the negative inputs, as we notice this in Fig. 9(a).

- Step1: START AReLU
- Step2: Import the required libraries.  
*keras, matplotlib, mnist, etc...*
- Step3: Loading the built-in MNIST Dataset.  
*mnist.load\_data()*
- Step4: Creating the model and add the layers: Input, Hidden and Output using dense layer from keras
- Step5: Compile the model by using the defined Loss Function, Optimizer and the metrics
- Step6: for i in range of epochs
  - 6.1 Fitting the Model by using the training set.
  - 6.2 Update AReLU parameter=new Loss value
- Step7: Evaluate the model on the testing set.
- Step8: STOP

Fig. 7. AReLU algorithm

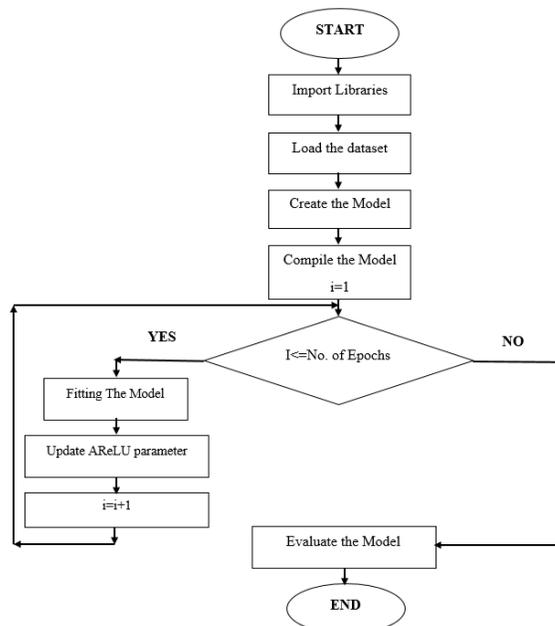


Fig. 8. AReLU flowchart

The used structure of the deep neural network is shown in Fig. 9 that composed of four layers, one input layer represents the input shape as 784 image pixels, two hidden layers each composed of 512 neurons and the final 10 neurons layer that characterize the output layer.

Binary Cross-Entropy Loss/Log Loss has been used as loss function in the model compilation process; where in this phase the loss function, the optimizer and the metrics are defined. This function is defined in Eq. (7); where N is the number of rows and M the number of classes.  $p_{ij}$  are the corrected probabilities, a negative average is used to compensate for negative values resulted from calculating log value of

corrected probabilities because their values range between 0 and 1. It is one of the most common loss functions used in multiclass classification problems. The value of this function decreases as the predicted probability converges to the actual label.

$$Loss = -\frac{1}{N} \sum_i^N \sum_j^M y_{ij} \log(p_{ij}) \quad (7)$$

Backpropagation in a network aims to make a change in the error value with respect to weights and this process is called derivative because its goal is to make a change in one value with respect to another. The first derivative of this function is:

$$\frac{d}{dx} f(x) = \frac{d}{dx} \begin{cases} \frac{d}{dx}(x), x > 0 \\ \frac{d}{dx}(l * x), x \leq 0 \end{cases}$$

$$\frac{d}{dx} f(x) = \frac{d}{dx} \begin{cases} 1, x > 0 \\ x, x \leq 0 \end{cases}$$

The function starts with any initial  $l$  value; say 0.1 and then it is automatically adapted according to the initial loss function value. Fig. 9(b) shows the Graphical Representation of AReLU Derivative. It is evident that the values of the derivative are close to zero but are not zero in the case of negative values.

The most effected activation function used in the output layer in the case of multi-layer classification problems is Softmax, which converts the raw outputs of a neural network into a vector of probability scores between 0 and 1. Its equations is defined in:

$$Softmax(o)_i = \frac{e^{o_i}}{\sum_{j=1}^N e^{o_j}}$$

Where  $\mathbf{o}$  is the input vector,  $e^{o_i}$  is the standard exponential function for  $o_i$ ,  $N$  is the number of classes in the multiclass classifier and  $e^{o_j}$  is the standard exponential function for output vector and  $e$  is the exponential which is equal nearly 2.718.

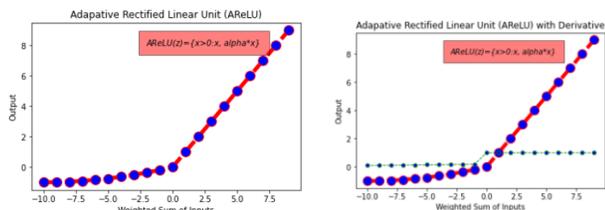


Fig. 9. AReLU graphical representation and its derivative

V. RESULTS

Fig. 10 illustrates the relationship between the training and validation accuracy over 20 epochs, the accuracy escalations are noticed in the first three epochs, indicating that the network is learning fast, thereafter the curve flattens, indicating that there is no need for more epochs to further training the model. The model accuracy was 98.8% (meaning

9880 of the 10000 images were predicted correctly!) and 120 images were wrongly tagged (1.2%).

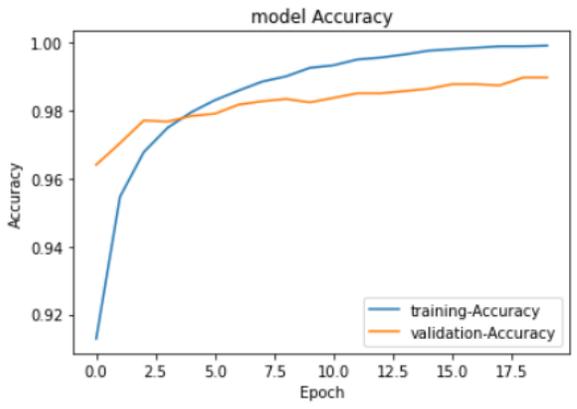


Fig. 10. AReLU model accuracy

AReLU gives better results based on the Misclassification Rate (MR) which is a measure of the percentage of observations that were incorrectly predicted by some classification model [21] and it's calculated as in

$$MR = \frac{Incorrect\ Predictions}{Total\ Predictions}$$

The MR for our model was 1.2% where for PReLU is 1.62 according to [22] as shown in Table I. This study measured the MR for different adaptive ReLUs including Sigmoid, tanh, MSAF, MSAF\_Symmetrical, ReLU, LReLU, ELU and adaptive tanh.

TABLE I. MR MEASUREMENTS FOR ACTIVATION FUNCTIONS ON MNIST DATASET

AF	MR
Sigmoid	7.01
Hyperbolic Tangent	1.86
MSAF	12.59
MSAF_Symmetrical	11.28
ReLU	2.08
LReLU	1.68
PReLU	1.6
ELU	1.88
Adaptive tanh	2.93

Reading Fig. 11 which illustrates the relationship between training and validation loss, we can see the rapid loss in the training set at the first two epochs while validation loss remained almost constant for several epochs, in contrast to the loss level of the training set, which means that the model can be generalized to unseen data.

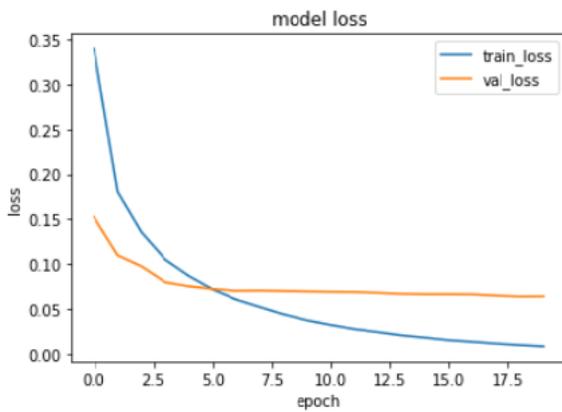


Fig. 11. AReLU model loss

## VI. CONCLUSION

This article produced automatic and adaptive activation function in which it retained inherent characteristics of ReLU with simplicity, high accuracy, speed and low loss ratio. Expected diminishing characteristic of Loss function value has exploited in implementing the AReLU, this function is used to measure the difference between the current output and the expected output. Cross-entropy type is used in developing the ReLU as one of the most widely used loss functions in machine learning due to its role in better generalization and faster model training. This function is used in binary and multi-class classification cases. AReLU is implemented by using Python programming language on MNIST dataset of handwritten digits to get 1.2% classification Rate. The model maintained the gains that the previous methods indicated, such as simplicity, low computational cost, no fixed coefficients, and adaptation in nature. In the future, AReLU will be applied to different data sets and work to reduce the rate of misclassification while maintaining the characteristics of simplicity, low computational cost, and no hyperparameters.

## ACKNOWLEDGMENT

The Author would like to express his gratitude to AlMaarefa University, Riyadh, Saudi Arabia, for providing funding to do this research.

## FUNDING

This research was funded by AlMaarefa University, Riyadh, Saudi Arabia

## REFERENCES

- [1] Nichols, James A., Hsien W. Herbert Chan, and Matthew AB Baker. "Machine learning: applications of artificial intelligence to imaging and diagnosis." *Biophysical reviews* 11.1 (2019): 111-118.
- [2] Sarker, Iqbal H. "Deep learning: a comprehensive overview on techniques, taxonomy, applications and research directions." *SN Computer Science* 2.6 (2021): 1-20.

- [3] Dubey, Shiv Ram, Satish Kumar Singh, and Bidyut Baran Chaudhuri. "Activation functions in deep learning: a comprehensive survey and benchmark." *Neurocomputing* (2022).
- [4] Fan, Jianqing, Cong Ma, and Yiqiao Zhong. "A selective overview of deep learning." *Statistical science: a review journal of the Institute of Mathematical Statistics* 36.2 (2021): 264.
- [5] Shaik, Nagoor Basha, et al. "A feed-forward back propagation neural network approach to predict the life condition of crude oil pipeline." *Processes* 8.6 (2020): 661.
- [6] Xie, Jingyi, and Sirui Li. "Training Neural Networks by Time-Fractional Gradient Descent." *Axioms* 11.10 (2022): 507.
- [7] Li, Yanyi, Jian Shi, and Yuping Li. "Real-Time Semantic Understanding and Segmentation of Urban Scenes for Vehicle Visual Sensors by Optimized DCNN Algorithm." *Applied Sciences* 12.15 (2022): 7811.
- [8] Chai, Enhui, et al. "An Efficient Asymmetric Nonlinear Activation Function for Deep Neural Networks." *Symmetry* 14.5 (2022): 1027.
- [9] Hahnloser, Richard HR, et al. "Digital selection and analogue amplification coexist in a cortex-inspired silicon circuit." *nature* 405.6789 (2000): 947-951.
- [10] Jarrett, Kevin, et al. "What is the best multi-stage architecture for object recognition?." *2009 IEEE 12th international conference on computer vision*. IEEE, 2009.
- [11] Nair, Vinod, and Geoffrey E. Hinton. "Rectified linear units improve restricted boltzmann machines." *Icml*. 2010.
- [12] Maas, Andrew L., Awni Y. Hannun, and Andrew Y. Ng. "Rectifier nonlinearities improve neural network acoustic models." *Proc. icml*. Vol. 30. No. 1. 2013.
- [13] He, Kaiming, et al. ". " *Proceedings of the IEEE international conference on computer vision*. 2015.
- [14] Clevert, Djork-Arné, Thomas Unterthiner, and Sepp Hochreiter. "Fast and accurate deep network learning by exponential linear units (elus)." *arXiv preprint arXiv:1511.07289* (2015).
- [15] Hajiabadi, Moein, et al. "Deep learning with loss ensembles for solar power prediction in smart cities." *Smart Cities* 3.3 (2020): 842-852.
- [16] Chadha, Ankita, Azween Abdullah, and Lorita Angeline. "A Comparative Performance of Optimizers and Tuning of Neural Networks for Spoof Detection Framework." *International Journal of Advanced Computer Science and Applications* 13.4 (2022).
- [17] Deng, Li. "The mnist database of handwritten digit images for machine learning research [best of the web]." *IEEE signal processing magazine* 29.6 (2012): 141-142.
- [18] Razak, H. A., et al. "Detection of Criminal Behavior at the Residential Unit based on Deep Convolutional Neural Network." *International Journal of Advanced Computer Science and Applications* 13.2 (2022).
- [19] Tan, Hong Hui, and King Hann Lim. "Vanishing gradient mitigation with deep learning neural network optimization." *2019 7th international conference on smart computing & communications (ICSCC)*. IEEE, 2019.
- [20] Cohen, Gregory, et al. "EMNIST: Extending MNIST to handwritten letters." *2017 international joint conference on neural networks (IJCNN)*. IEEE, 2017.
- [21] Maach, Anas, et al. "An Intelligent Decision Support Ensemble Voting Model for Coronary Artery Disease Prediction in Smart Healthcare Monitoring Environments." *arXiv preprint arXiv:2210.14906* (2022).
- [22] M. M. Lau and K. Hann Lim, "Review of Adaptive Activation Function in Deep Neural Network," *2018 IEEE-EMBS Conference on Biomedical Engineering and Sciences (IECBES)*, 2018, pp. 686-690, doi: 10.1109/IECBES.2018.8626714.

# Espousing AI to Enhance Cost-Efficient and Immersive Experience for Human Computer Interaction

Deepak Chaturvedi<sup>1</sup>, Ashima Arya<sup>2</sup>, Mohammad Zubair Khan\*<sup>3</sup>, Eman Aljohani<sup>4</sup>, Liyakathunisa<sup>5</sup>, Vaishali Arya<sup>6</sup>,  
Namrata Sukhija<sup>7</sup>, Prakash Srivastava<sup>8</sup>

Department of Computer Science and Engineering, KIET Group of Institutions, Delhi-NCR, Ghaziabad, India<sup>1,2</sup>

Department of Computer Science and Information, Taibah University, Medina, Saudi Arabia<sup>3</sup>

Department of Computer Science-College of Computer Science and Engineering, Taibah University, Madinah, Saudi Arabia<sup>4,5</sup>

Department of Computer Science and Engineering, GD Goenka University, Gurugram, India<sup>6</sup>

Department of Computer Science and Engineering, Banasthali Vidyapith, Rajasthan, India<sup>7</sup>

Department of Computer Science and Engineering, Graphic Era (Deemed to Be University), Dehradun, India<sup>8</sup>

**Abstract**—Because of recent technological and interface advancements in the field, the virtual reality (VR) movement has entered a new era. Mobility is one of the most crucial behaviours in virtual reality. In this research, popular virtual reality mobility systems are compared, and it is shown that gesture control is a key technology for allowing distinctive virtual world communication paradigms. Gesture based movements are very beneficial when there are a lot of spatial restrictions. With a focus on cost-effectiveness, the current study introduces a gesture-based virtual movement (GVM) system that eradicates the obligation for expensive hardware/controllers for virtual world mobility (i.e., walk/ jump/ hold for this research) using artificial intelligence (AI). Additionally, the GVM aims to prevent users from becoming dizzy by allowing them to change the trajectory by simply turning their head in the intended direction. The GVM was assessed on its interpreted realism, presence, and spatial drift in the actual environment in comparison to the state-of-the-art techniques. The results demonstrated how the GVM outperformed the prevailing methodologies in a number of common interaction components. Additionally, the empirical analysis showed that GVM offers customers a real-time experience with a latency of ~65 milliseconds.

**Keywords**—Artificial intelligence; dizziness; gestures; human computer interaction; user experience; virtual reality

## I. INTRODUCTION

Virtual reality (VR) has been around for years, but it has only lately piqued the attention of customers and professionals as the technology grows increasingly economically viable. VR economies are exploding, with the overall global demand estimated to approach four billion revenues by 2025, involving 45 million VR headset deployed and a global population coverage of 3% [1]. Human-computer interaction approaches in the earlier years compelled human behaviour to conform to the computer's capabilities; however, VR perspective is unique in that the computer now must mirror the actual environment to deliver the most authentic view feasible. To provide individuals an immersive experience, VR develops a variety of participation activities relating to visual, auditory, and tactile sensitivities. Widely available methodologies for

VR mobility are heavily reliant on a controller to explore and move, or actual relocating in a constrained geographic space, disregarding proliferating necessities on the strategy for travelling an unregulated virtual space by physically strolling the user's legs, which causes fatigue [2]. The most extensively used VR movement methods are listed below.

1) *Gadgets*: A frequent strategy for navigation in the VR world is to use gadgets such as joysticks and head orientation tracking with Gyro in VR head mounted displays. For consumers focused on control movement in VR, these gadgets are intuitive and comfortable, straightforward to use, and productive. However, because of a perceptual mismatch [3] between visual and vestibular inputs [4], joysticks frequently influence the vision to act swiftly [5] and erratically, creating dizziness [6].

2) *Teleportation*: Another typical strategy for reducing dizziness is to provide many gateway locations allowing players to swiftly move from one location to the next. Unfortunately, due to the discontinuous movement that negatively impacts the user's experience and may induce vertigo, these tactics are not organic enough to boost the interactive experience in the virtual environment [7].

3) *Walking-in-place (WIP)*: The WIP approach allows users to travel in a specific location while controlling the character's motion and orientation using real body gesture detection sensors such as Microsoft Kinect [8]. This technique enhances the matching among mechanoreceptors of data from a person's body movements and tactile senses through machine screens, rendering it more natural and potentially lowering operator dizziness. Nevertheless, this technique requires the user to remain in one place and use their entire body, as well as a large amount of underlying hardware, that are costly and not available to all. A good travel experience, on the other hand, must cause less fatigue in a walk-through arrangement [9].

4) *Hand gestures*: In virtual reality, a gesture is a stance or motion of the user's body which is employed as input. The

\*Corresponding Author

WIP approaches tend to depend on the same gesture for triggering forward perspective motion: leg gestures like those used while climbing a stairway [10]. This motion emerges to be more exhausting than actual walking. Hand gestures, on the other contrary, can be an organic and efficient technique for controlling motions in virtual reality. Furthermore, movement based on hand gestures has the benefit of requiring less exertion and decreasing dizziness as it can be performed while sitting or standing [11]. For VR engagement, there are a range of gesture communication devices that facilitate communication more authentically with items in the virtual environment. These gesture communication devices are classified on the mode of input as wearable sensor, touch, and computer vision [10].

The authors in current research introduce a gesture-based virtual movement (GVM) system to facilitate an inexpensive solution for supporting individuals with walk-through activities in virtual worlds, which allows customers to unwind while sitting or standing in a place as if they're in reality

#### A. Key Contributions

The authors' goal in this study is to enrich the user's immersive experience. The following are the major findings of this research.

1) *Cost efficiency*: GVM is a low-cost solution that eliminates the requirement for any additional costly gesture recognition gear.

2) *Purging dizziness*: GVM reduces dizziness by letting users modify their trajectory by merely tilting their head in the desired direction and hand movements for gesture recognition to move in VR.

3) *Handling strain*: GVM relieves the user of physical strain.

4) *Usability*: The usefulness of the suggested approach is demonstrated by user input on several factors such as interpreted realism, presence, and spatial drift in the real world.

5) *Real-time experience*: With a latency of ~65 milliseconds, the suggested system offers consumers a real-time experience.

#### B. Paper Organization

The manuscript is further divided into sections. Section II presents a brief literature survey of the various VR techniques. Section III introduces the proposed model, GVM. Section IV explains the experimentation done and the results achieved that highlights the suitability of GVM. Section V concludes this research. Finally, Section VI highlights the future work.

## II. LITERATURE SURVEY

Table I shows a comparison of various widely used VR movement methods based on the dimensions of motion sickness and physical strain. Hand gestures have been shown to be a remedy for motion sickness and physical strain; however, using a hand gesture detection system necessitates the acquisition of expensive gears. Thus, the authors introduce a GVM system to facilitate an inexpensive solution for

supporting individuals with walk-through activities in virtual worlds, which allows customers to unwind while sitting or standing in a place as if they're in reality.

TABLE I. COMPARISON OF MOST EXTENSIVELY USED VR MOVEMENT METHODS

#	Methodology	Motion Sickness	Physically Straining
1	Gadgets	Yes	Yes
2	Teleportation	Yes	No
3	Walking-in-place	No	Yes
4	Hand Gestures	No	No

Mine [12] proposes using hand-based communication to manage mobility and walk-through in a simulated world. An elevated hand-gesture tracer gadget, such as Leap Motion, is a unique technology which delivers input via hand gesture mapping, allowing for bare-hand interactivity [13] in a three-dimensional world. Ni et al. [14] investigates menu selection employing freehand signals, whereas Kulshreshth et al. [15] provides the findings of the first thorough research on finger-count panels to assess their suitability for 3D menu choice applications. Beattie et al. [16] demonstrates a CAD Engagement Facility that allows users to deconstruct a kinematic model in virtual reality and operate and analyse constituent parts. Lee et al. [17] offer TranSection, a hand-based communication strategy for executing a strategy game in virtual reality. Salomoni et al. [18] describes research in which recreational virtual world interfaces are reconsidered in view of the rise of head-mounted displays. These concepts, unfortunately, do not yet include how to handle walk-through activity in a simulated world.

Numerous studies have investigated ways to execute a natural and pleasant interaction approach in VR employing Leap Motion to solve this research gap. Codd-Downey et al. [19], for instance, offers a finger tracking movement approach that uses a 2DOF driving paradigm like typical mouse and keyboard control in 3D computer gaming. Khundam [20] presents a novel engaging single-hand-gesture control drive system with palm norm. The results reveal that controlling tour activity with hand gestures is more natural than to use a joystick. There are several aspects of VR controller hardware for diverse approaches, and some studies have developed a system that gathers multiple devices for a certain objective. The Oculus Rift and Leap Motion have lately been employed in several studies, particularly in virtual reality. Programmers are particularly interested in studying usage patterns and determining what the most productive utility for them in the future is through VR engagement.

Prior studies on in-air controllers and hand monitoring intended to develop and deploy VR applications. The precision of hand monitoring is critical for a reliable system. Sato et al. [21] provide a technique for monitoring a user's hand in three dimensions and identifying hand gestures in real time even without any intrusive sensors connected to the hand. Several cams are used to assess the location and direction of a user's hand floating in 3D environment. A neural network that is adequately trained recognises specified motions in a rapid and reliable fashion. 3D item processing for a desktop machine and

3D movement for a big holistic projection system are two typical applications. Many studies have been done on hand gestures and their uses. Chastine et al. [22] describe research comparing single hand gestures to typical keyboard, mouse, and controller input of first-person gameplay. The purpose of this study is to enable game analysts, architects, and builders to better understand how to include gesture control in current applications. The findings demonstrate that in FPS games, human rehearsals are crucial for gestural-based gaming system performance. As people continued through the activities, users were increasingly skilled at using the gadget, indicating that gesture-based handling can be used by users with no prior knowledge. This feedback helps programmers to employ Leap Motion as a device in virtual reality and ensures that there is a compelling incentive for them to do so in the long term.

Many people use virtual reality headsets to interact with 3D models. Stefan Greuter and David J. Robert [23] present the SpaceWalk technology. This system, which consists of two hardware devices: a motion sensing unit and a cordless VR gear, allows for low-weight full-body VR experience while wandering around the living area. The preponderance of the equipment in this system are made up of an Oculus Rift (DK1) HMD and a backpack tablet which operates standard VR program (Unity3D) alongside their extension script that connects all the elements. Participants may move and engage with things in the virtual world in this research's living area, however this framework is not designed for huge VR environments. Weibel et al. [24] describe how to build a moderate, fully interactive, stochastic virtual world setup that allows users to naturally perceive intangible cultural assets. They look at new technology including the Oculus Rift virtual reality headset, Microsoft Kinect, and the Leap Motion controllers. When it comes to constructing HMD VR situations, modern technologies such as the Oculus Rift HMD, Microsoft Kinect, and Leap Motion provide excellent results.

The usage of the Kinect or Leap Motion in conjunction with organic conversational inputs lets users engage directly with the virtualized world. However, because of the user's movement control, this VR system is generally limited to the comparatively small region in front of the sensing element. As a result, adopting engaging hand gestures for motion in VR will increase the VR system's admin tools via rigorous positioning and replacing previous techniques.

Users can employ an expanding number of input gadgets to engage with systems and apps. When building applications for technological innovations, though, there are no defined interface guidelines or benchmarks, and the customer satisfaction suffers the consequences. Jake Araullo and Leigh Ellen Potter [25] give a study that investigates the perspectives of a set of people who used the Oculus Rift and the Leap Motion device to play. The incorporation of blended conventional and non-traditional input methods, as well as depending on existing interface paradigms when leveraging innovative methods, were found to have a detrimental impact on system adoption in this study.

The present research proposes a gesture-based virtual movement (GVM) system that eliminates the need for pricey equipment for immersive virtual movement (i.e., walk/jump/hold for this research) with a focus on affordability. By enabling users to alter the trajectory by merely rotating their head in the desired direction, the GVM also seeks to prevent users from feeling dizzy.

### III. PROPOSED MODEL

The goal of authors is to employ user hand gestures to create movement in the virtual environment. The suggested GVM's overall process flow is shown in Fig. 1. The overall procedure is segmented into the following:

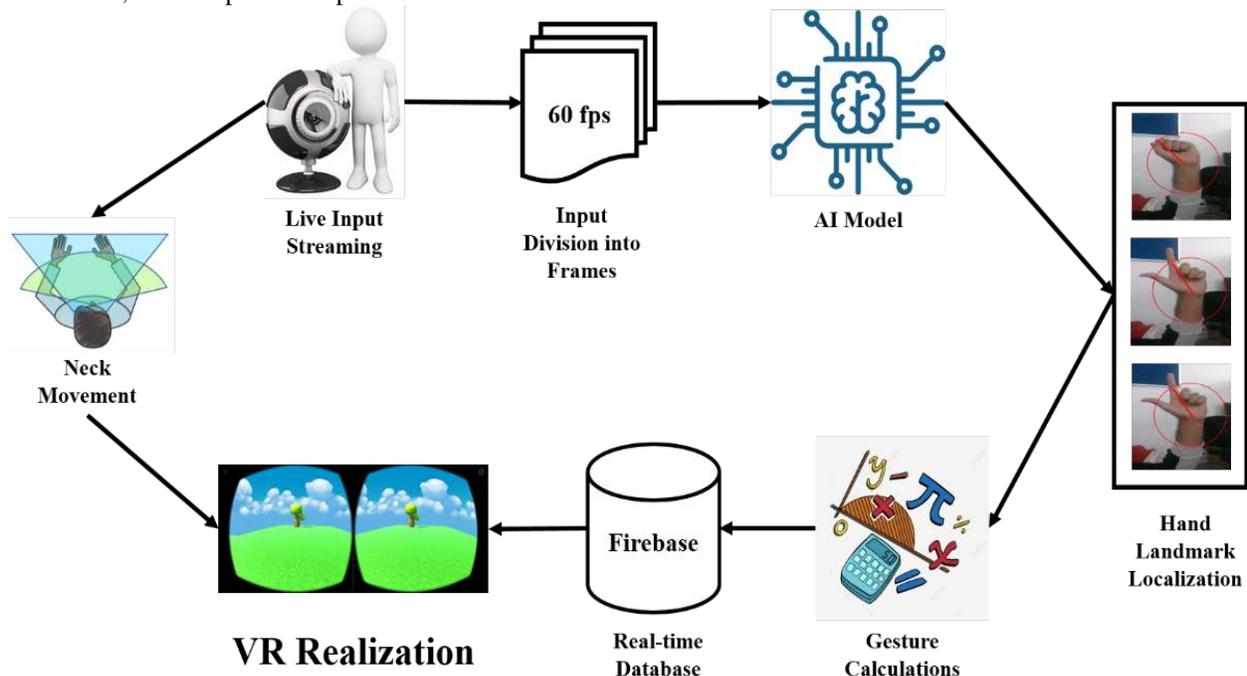


Fig. 1. Process flow of the proposed model

- Input Processing
- Gesture Identification
- Database Interaction
- VR Realization

### A. Input Processing

The suggested approach uses streamed live video as an input. The footage is divided into frames at a 60 frames per second rate rather than being supplied directly to the model. Webcam data is used to provide each frame to the model that are further fed to an AI model for recognizing the hand landmarks.

### B. Gesture Identification

An artificial intelligence (AI) model built on top of MediaPipe's [26] recognises the hand motions. A platform for creating pipelines that do interpretation over any type of sensory input is entitled MediaPipe. The AI model works in two phases i.e., palm detection and hand land-marking achieved through a palm detector and hand landmark model, respectively.

- Using an aligned hand bounding box, a palm detector identifies palms on a whole input picture. A single-shot detector model tailored for cellular real-time is utilised to find the first hand placements.
- A hand landmark model which generates high-definition 2.5D landmarks based on the palm detector's clipped hand bounding box. After detecting the palm across the entire picture, a second hand landmark model uses regression, or direct location projection, to carry out exact feature point placement of 21, 3D hand-knuckle positions inside the identified hand areas. Only six of them were used for the suggested gestures model, as seen in Fig. 2. The model acquires a reliable inherent hand posture depiction and is unaffected by self-occlusions or semi-transparent hands. In every instance, the landmarks are almost perfectly spotted.

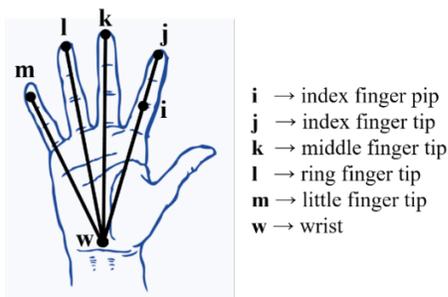


Fig. 2. 3D hand-knuckle coordinates used in proposed model

Rather than employing hard computing, authors have opted to soft computing to identify gestures more precisely. The proportionate placements of various landmarks serve as the basis for codes. Utilizing relative locations, authors programmed three distinct gestures: hold, move, and jump.

- Hold: In this gesture, the radius is formed by the line connecting the wrist and the tip of the index finger, as shown in Fig. 3(a).
- Jump: In this gesture, the radius of the circle is established by the line connecting the wrist and the pip of the index finger, and the tips of the remaining fingers are contained within the circle, as seen in Fig. 3(b).
- Move: In this motion, all finger tips are located outside the circle, with the radius being the line between the wrist and the tip of the index finger (see Fig. 3(c)).

Every live streamed hand gesture is labelled as either one of three (hold/jump/move) and further the gesture calculations are used to classify the gestures accurately. The gestures calculation starts with the identification of Euclidian distance,  $\delta$  between the coordinates of wrist ( $x_w, y_w$ ) and index finger pip ( $x_i, y_i$ ) as per equation (1)

$$\delta = \sqrt{(x_w - x_i)^2 + (y_w - y_i)^2} \quad (1)$$

After the  $\delta$  is calculated, the behavior of index (j), middle (k), ring (l) and pinky (m) fingers are identified using equation (2) and (3)

$$a = \Psi_{finger \in j}(x_{finger}^2 + y_{finger}^2 - \delta^2) \quad (2)$$

$$b = \Psi_{finger \in k,l,m}(x_{finger}^2 + y_{finger}^2 - \delta^2) \quad (3)$$

The gesture,  $\Omega$  is the calculated based on equation (4)

$$\Omega = \begin{cases} \text{hold, if } a < 0 \text{ and } b < 0 \\ \text{jump, if } a < 0 \text{ and } b > 0 \\ \text{move, if } a > 0 \text{ and } b > 0 \end{cases} \quad (4)$$

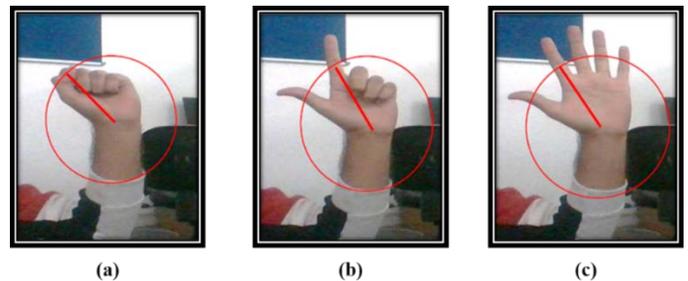


Fig. 3. Gestures for (a) Hold; (b) Jump; and (c) Move

### C. Database Interaction

The AI Model [27, 28] subsequently sends the gesture to the real-time database (Firebase in current research), which updates the motion parameter with the potential movement gestures (isMove, isHold or isJump). The Firebase database gives the system the most recent value of the information as well as modifications to that information by using a single API. The clients are able to retrieve their data from any platform, including the web and mobile devices, owing to real-time synchronization.

On the other side, the Unity3D Engine [29] is coupled to the real-time database. The Unity3D engine serves as the base layer for the present VR experience. Additionally, C# scripting is used to fetch data from the real-time database each time a database update is triggered.

#### D. VR Realization

The user has complete freedom to roam around the area and may utilise gestures to commence any movement. In the virtual environment, neck movement provides the directional input. Users of Virtual Reality (VR) may freely spin their heads 160° while viewing the surroundings owing to rotational tracking (as presented in Fig. 4).

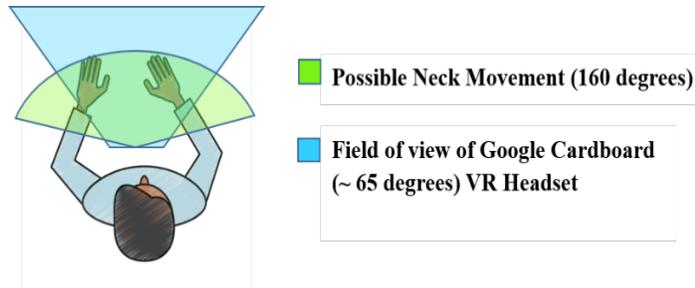


Fig. 4. Neck movement for VR environment

The user's head movement determines how the player rotates. The trajectory of the avatar's movement is controlled by the Head Mounted Display (HMD) spin. The field of view (FOV) of cell phone based VR headsets like Google Cardboard is only approximately 65 degrees (as presented in Fig. 4). The avatar travels both in translation and rotation inside the VR environment based on the data received.

### IV. EXPERIMENTS AND RESULTS

Current research's objective is to assess the system behavior, empirical characteristics, and experience aspects that are most important for VR locomotion [30]. To evaluate the efficacy of GVM, a comparison research using four approaches i.e., walking-in-place, controller/joystick, teleportation, and GVM (the proposed approach) is conducted. Current research investigates the propose model on two aspects i.e., 1) Latency and 2) User Experience.

#### A. Environmental Setup

The HTC Vive headgear and Epic Games' Steam VR SDK for Unreal Engine 4 were used in the development of the experimented-with VR locomotion methods. With a display resolution of 1080 x 1200 (2160 x 1200 combined pixels), 90 Hz refresh rate, 110 field-of-view, and complete 360 room-scale human monitoring, the HTC Vive headgear allows high-fidelity visuals. It is well known in the commercial VR industry and is made to use room-scale equipment, which uses sensors to transform a place into a 3D world. The HTC Vive monitoring system, an extra sensor that can be utilized to monitor tangible goods and translate them into activities or items in the simulated space, is supported by the system. Using a pristine HD 720p/30 fps camera with a diagonal field of view of 55 degrees and automatic light adjustment, the Hand Gesture Detection feature of Logitech C270 Digital HD webcam is employed.

1) *Walking in place*: The participant's limb motions during walking in place must be converted into virtual reality activity. The participants' right foot-mounted HTC Vive tracker and HTC Vive controllers were used to record and, respectively, manage the VR movement velocity and

direction. The VR movement velocity is closely correlated with the users' actual walking speed; that is, the quicker the participants moved around in actual situations, the quicker their avatars moved in the simulated space. Right footstep speed is used to imitate left footstep speed. The HTC Vive controllers' orientation affected the motion direction. Users have to manually turn themselves in the intended way in order to adjust the movement's trajectory.

2) *Controller/joystick*: In this approach, the type of controller can be anything from a straightforward joystick to a gaming remote or a keyboard. To enable controller-based VR movement, the HTC Vive controllers have been used as a touchpad. Motion is initiated by tapping the touchpad, and the velocity of motion is controlled by where the thumb is placed on the touchpad. The HMD system displayed a directional line to indicate the direction of motion, which has been governed by the orientation of the HTC Vive controllers.

3) *Teleportation*: With this method, you may point or use a controller to indicate where you want to teleport to. The HTC Vive Controllers' grip trigger is used. Whenever the trigger is pulled, a graphical signal that showed the movement's location, a ray accompanied by a marking on the simulated ecosystem's ground appeared. The trigger is pushed to initiate movement. The teleportation's orientation has been decided by the participant's body orientation.

4) *GVM*: GVM is a low-cost solution that eliminates the requirement for any additional costly gesture recognition gear. It reduces dizziness by letting users to modify their trajectory by merely tilting their head in the desired direction and hand movements for gesture recognition to move in VR. GVM relieves the participant from the physical strain. The usefulness of the suggested approach is demonstrated by user input on several factors such as interpreted realism, presence, and spatial drift in the real world. With a latency of approximately 65 milliseconds, the suggested system offers consumers with a real-time experience.

The virtual 3D environment is build using Unity 3D Engine, version 2019.4.40f1 (LTS) and deployed for Android and IOS platform. A simple Unity3D scene as presented in Fig. 5 is setup for the survey having 3D assets and paths to explore.

Participants can perform various movement actions like Jump/Hold/Move within the environment and move freely. This investigation gathers information to create an assessment of the strategies' efficacy in real-world settings. It moreover gathers information through semi-structured questionnaires to create a "rich description" of the perspectives of the participants.

#### B. Latency

The duration that it takes for information which is fed at an end of the connection to appear at the opposite end is referred to as latency. Typically, authors gauge how long it takes for information to go from one end to the other. In this setup we actually measure the round trip time (RTT), the "latency" (time of event from real-time database to Unity3D Engine) can easily be estimated as  $\Delta_i = 0.5 * RTT$ , where  $\Delta_i$  represents the latency.

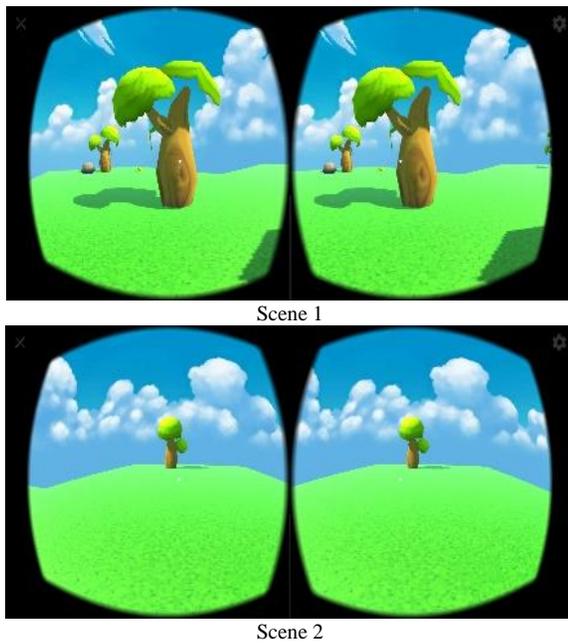


Fig. 5. Virtual environment created for experimentation

The tests were run with both/all clients on the same machine, located behind a 100mbits connection. For the Firebase Real-time Database, the location [us-central1] was selected. The average latency for three hundred observations is used to summarize the time offset between the data flow and RTT is calculated as approximately 65 milliseconds.

### C. User Experience

Utilizing the Game Experience Questionnaire (GEQ) [31], the user experience is assessed. Due to its capacity to address a broad variety of experiential aspects with strong reliability, the GEQ is a customer experience questionnaire which has been utilised in numerous areas (including gaming, virtual reality, and location-based services) [32, 33]. In numerous research on subjects including VR education [34], haptic engagement in VR [35], virtual reality orientation and mobility [36, 37] and virtual reality entertainment [38], the usage of GEQ has also been validated in the VR arena. The GEQ's Competence, Sensory and Imaginative Immersion, Flow, Tension, Challenge, Negative Affect, Positive Affect, and Tiredness categories are deemed pertinent and helpful for current investigation of the underlying strategies. According to a sequence of phrases in the GEQ questionnaire, the participant has been prompted to describe how he or she experienced throughout the encounter. It had 16 assertions that have been scored on a five-point severity scale from 0 ('not really') to 4 ('strongly') and included phrases like "I forgot everything around me". At the beginning of the research, demographic information was gathered, including age, gender, regularity of VR exposure ('never, seldom, often, and every day'), and familiarity with VR technology.

1) *Analysis of participants:* Within our institution zone, the participants have been sought for between October 2022 and December 2022. Participants needed to be physically capable of using VR technology, although prior VR

experience wasn't really necessary. Participants have been informed of the possibility for dizziness as well as their right to withdraw from the research at any moment. To be a part of the study, every participant provided their informed permission.

The four VR locomotion strategies were tested on thirty people (N = 30, mean age: 22.7, male/female: 18/12). Twelve individuals had only sometimes used virtual reality (VR), whereas eight people had used it regularly. Ten participants had never utilized VR. Twenty participants had earlier used VR; six had done so with HMDs and portable VR headsets, eleven had done so solely with HMDs, and three had done so only with portable VR headsets. Each participant finished the episode satisfactorily.

2) *Methodology:* After providing the informed consent, the participants responded to demographic and VR encounter forms within approximately ten minutes' duration. Then, the participants had additional trial opportunities to discover at their leisure and witness a "clean" rendition of the VR world, that is, one that had no time restrictions and did not use the VR locomotion approach for an average five minutes). The exercise was then completed by the participants within a duration of ten minutes on average. The participants may provide vocal comments while traversing using the GVM approach, and the investigators have been taking notes in order to tackle these issues in the discussion. The GEQ questions have been completed once the work has been finished within a duration of five minutes on average.

3) *Results & discussions:* There have been thirty tasks in total, one for each participant. The typical assignment took thirty-seven minutes to complete. The GVM technique stood out magnificently outstanding in the majority of the GEQ constituents (i.e., Competence, Sensory and Imaginative Immersion, Flow, Tension, Challenge, Negative Affect, Positive Affect, Tiredness) after the couple mean leader board analysis (depicted in Table II).

TABLE II. GEQ RANKING FOR VR LOCOMOTION TECHNIQUES

S.No.	VR Locomotion Techniques	Benchmark GEQ Rating
1	WIP	Ok
2	Controller	Borderline excellent
3	Teleportation	Good
4	GVM	Excellent

The participants felt that WIP offered excellent degrees of immersion because of its authentic and organic movement. However, most participants indicated the approach to be exhausting due to the difficulty of translating actual body action to virtual reality motion. Others said this function added a certain amount of amusement, enjoyment, and exercise. Eventually, amidst the investigators attempting to take all necessary precautions, such as setting up a virtual boundary structure and an open area, participants still reported experiencing a pause in their exploration in the simulated space due to their fear of running into actual physical items in the real

world. It was discovered that the controller/joystick VR movement was simple to use and has been described as "pleasant," "simple," and "pleasant." However, a few users mentioned experiencing brief motion nausea at the beginning of the questionnaire job.

Owing to its visible 'jumps' and irregular mobility, teleportation was deemed the weakest engaging of the four modalities. On the contrary, the participants judged GVM to be more engaging and competent than the majority of the GEQ aspects. In addition to reducing fatigue compared to WIP, it also eliminated motion sickness brought on by hand gestures. The user had fewer difficulties using GVM because of the predetermined hand movements. The majority of participants praised GVM and rated it as the easiest and perhaps most enjoyable approach. Fig. 6 and Table III shows a normalized mean value (NMV) between 0 - 10 of each technique for every GEQ Component.

- **Competence:** The tests revealed that the Competence scores for the various strategies differed statistically significantly. On analyzing the Competence grade of GVM with all other techniques, the NMV showed substantial variances, favouring GVM with a NMV of 8.87.
- **Sensory and Imaginative Immersion:** The findings indicated statistically significant contrasts among the four strategies for the Sensory and Imaginative Immersion aspect favouring GVM in close vicinity to WIP.

- **Flow:** Following the test, there have been no appreciable variations in the Flow component amongst the four strategies; nonetheless, participants gave the Teleportation approach a higher rating.
- **Tension:** GVM obtained the lowest mean value (3.64) in the assessment, followed by WIP, Teleportation, and Controller, in that order.
- **Challenge:** The challenge score between GVM and Teleportation differed significantly according to the MSR, showing that GVM (mean value: 5.02) is a less difficult approach than teleportation and others.
- **Negative effect:** The testing activity for Negative effect demonstrated significant differences throughout all technique analyses. In the mean assessment of GVM, WIP, Controller, and Teleportation, the GVM showed minimal negative effect. Teleportation on the other hand put extra strain on the participants due to continuous transition in the virtual world.
- **Positive effect:** The GVM and teleportation had greater value, but the assessment did not reveal any changes in the Positive Affect component amongst the four approaches.
- **Tiredness:** The Tiredness aspect exhibited substantial variations according to the assessment. In each instance, GVM, Teleportation, and Controller all scored much lower on Tiredness compared to WIP

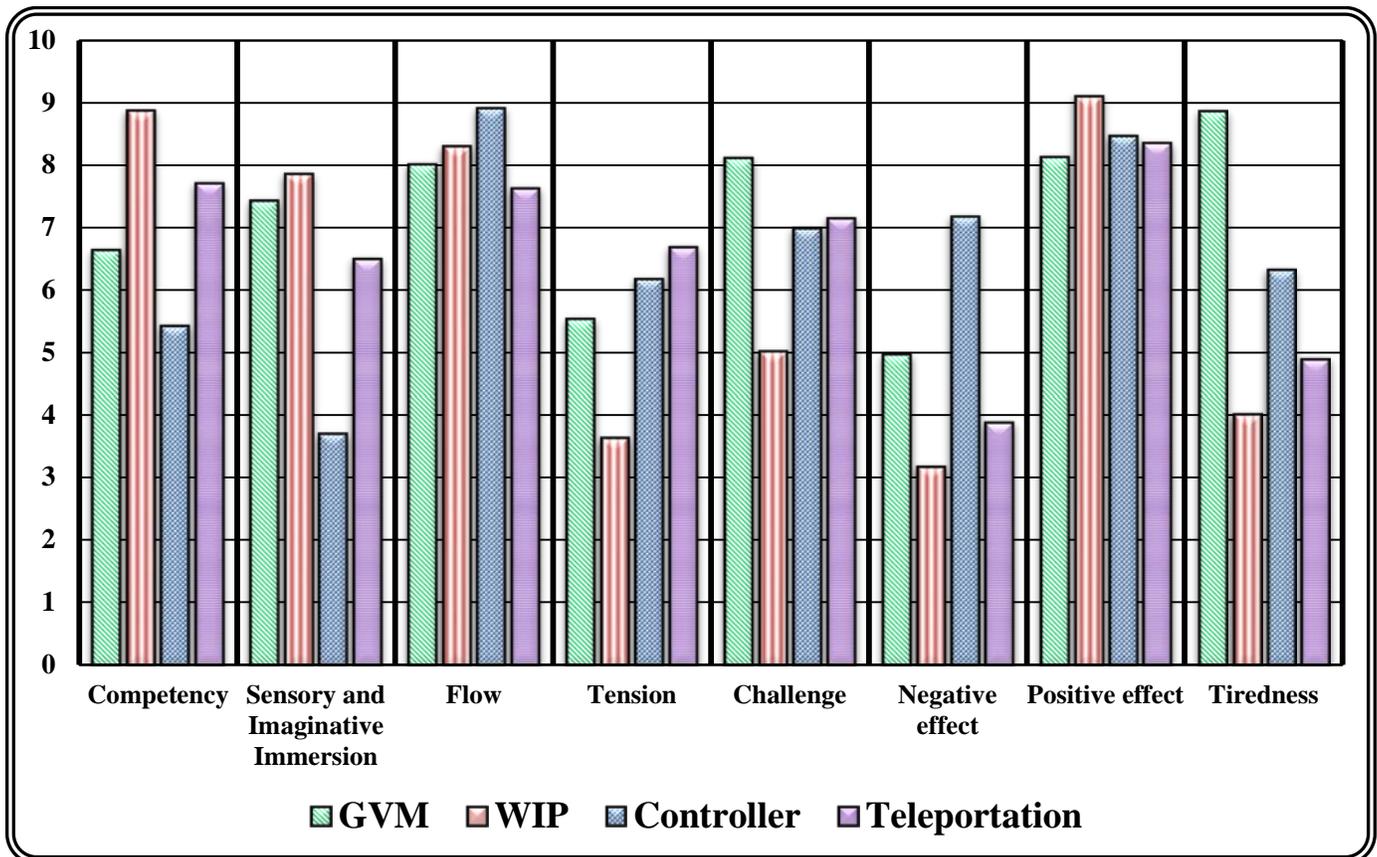


Fig. 6. Normalized mean value for various GEQ components

TABLE III. GEQ COMPONENTS NORMALIZED MEAN VALUE FOR EXISTING TECHNOLOGIES

Evaluation Parameter	Proposed Method (GVM)	WI P	Controll er	Teleportat ion
Competency	6.63	8.87	5.42	7.71
Sensory and Imaginative Immersion	7.42	7.86	3.7	6.5
Flow	8	8.3	8.9	7.63
Tension	5.53	3.64	6.17	6.69
Challenge	8.1	5.02	6.98	7.15
Negative effect	4.96	3.18	7.17	3.89
Positive effect	8.12	9.1	8.46	8.35
Tiredness	8.85	4.02	6.32	4.9

Fig. 7 shows overall positive components mean values for each techniques, positive GEQ components include Competence, Sensory and Imaginative Immersion, Flow, Positive Effect. In this the mean value of GVM is highest. Further, Fig. 8 shows overall negative components mean values for each technique, negative GEQ components include Tension, Challenge, Negative Effect, Tiredness. In this the mean value of GVM is lowest.

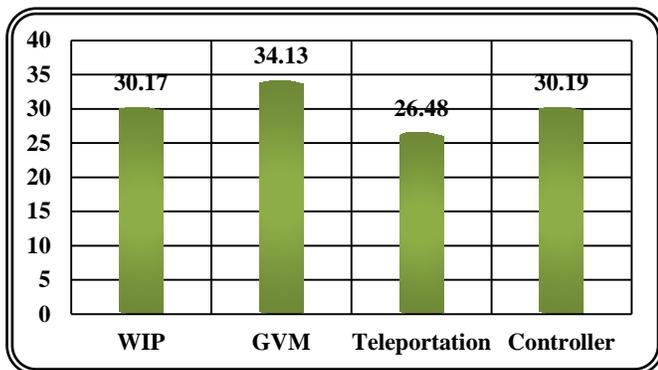


Fig. 7. Mean value for various positive GEQ components

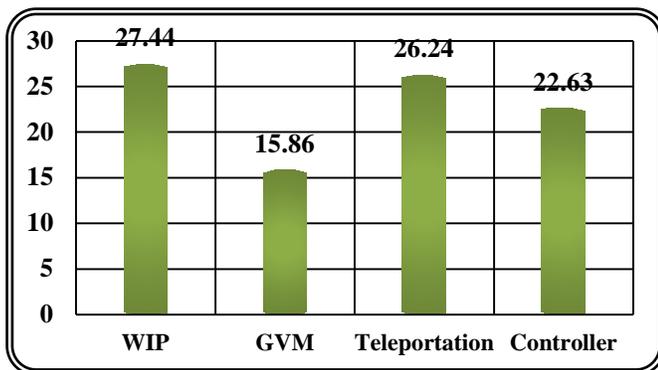


Fig. 8. Mean value for various negative GEQ components

### V. CONCLUSIONS

A range of hand gestures must be recognised and reliably classified by gesture recognizers in order to provide improved

user interfaces for Virtual or Augmented Reality goods. This study compares the most common virtual reality mobility systems and finds that gesture control is a crucial technology for enabling unique virtual world communication paradigms. The present research proposes a gesture-based virtual movement (GVM) system using artificial intelligence (AI) that eliminates the need for pricey equipment for immersive virtual movement (i.e., walk/jump/hold for this research) with a focus on affordability. By enabling users to alter the trajectory by merely rotating their head in the desired direction, the GVM also seeks to prevent users from feeling dizzy. In comparison to cutting-edge methods, the GVM's interpreted realism, presence, and spatial drift in the real world were evaluated. According to the empirical analysis, GVM offers customers a real-time experience with a latency of ~65 milliseconds. Additionally, the results demonstrated how the GVM outperforms the existing techniques in many standard interaction elements.

### VI. FUTURE WORK

A proof-of-concept for using hand motions identified by computer vision to enable movement in a virtual world is provided by the work discussed in this paper. However, there remains lots of opportunities for enhancement and more research.

- *Enhancement of Gesture Lexical Items:* Authors aim to increase the number of hand gestures available for use in directing the movement of the virtual world. This can entail introducing fresh motions that let users navigate across the area or control items.
- *User Interface Layout:* It will be crucial to create a user interface that is simple to understand and use as the system grows increasingly complicated and feature-rich. We will investigate several methods for creating a user interface that permits individuals to swiftly and simply manipulate the virtual world in upcoming work.
- *User Experience:* In order to enhance the user experience, we will research several ways to let users know when their gestures have been effectively identified. This can entail adding haptic or visual feedback features to let users know when they have performed a motion correctly.
- *Reduce delay:* Making the virtual reality environment feel more realistic by lowering system delay may significantly enhance user experience. Optimizing data communication between the AI model and the Unity3D engine is one method for lowering latency. This can entail transferring data using more effective methods or requiring less data to be transferred in real-time.

In conclusion, there are a wide range of prospective directions for further research in this field, including increasing the gesture lexicon, improving gesture detection, enhancing user interface, offering users input, and lowering latency. Future research in these areas has the potential to dramatically improve user engagement and increase the effectiveness and efficiency of the gesture-based movement mechanism.

REFERENCES

- [1] "2020 in review: Virtual reality gets real | Industry Trends | IBC," 10 December 2020. [Online]. Available: <https://www.ibc.org/trends/2020-in-review-virtual-reality-gets-real/7106.article>. [Accessed 23 July 2021].
- [2] W. Kim and S. Xiong, "User-defined walking-in-place gestures for VR locomotion," *International Journal of Human-Computer Studies*, vol. 152, p. 102648, 2021.
- [3] S. S. Chance, F. Gaunet, A. C. Beall and J. M. Loomis, "Locomotion mode affects the updating of objects encountered during travel: The contribution of vestibular and proprioceptive inputs to path integration," *Presence*, vol. 7, no. 2, pp. 168-178, 1998.
- [4] B. K. Jaeger and R. R. Mourant, "Comparison of simulator sickness using static and dynamic walking simulators," in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 2001.
- [5] S. Batra, R. Khurana, M. Z. Khan, W. Boulila, A. Koubaa and P. Srivastava, "A Pragmatic Ensemble Strategy for Missing Values Imputation in Health Records," *Entropy*, vol. 24, no. 4, p. 533, 2024.
- [6] K. M. Stanney, R. S. Kennedy, J. M. Drexler and D. L. Harm, "Motion sickness and proprioceptive aftereffects following virtual environment exposure," *Applied ergonomics*, vol. 30, no. 1, pp. 27-38, 1999.
- [7] D. A. Bowman, D. Koller and L. F. Hodges, "Travel in immersive virtual environments: An evaluation of viewpoint motion control techniques," in *Proceedings of IEEE 1997 Annual International Symposium on Virtual Reality*, 1997.
- [8] P. T. Wilson, K. Nguyen, A. Harris and B. Williams, "Walking in place using the Microsoft Kinect to explore a large VE," in *Proceedings of the 13th ACM SIGGRAPH International Conference on Virtual-Reality Continuum and its Applications in Industry*, 2014.
- [9] Z. Yan, R. W. Lindeman and A. Dey, "Let your fingers do the walking: A unified approach for efficient short-, medium-, and long-distance travel in VR," in *2016 IEEE symposium on 3D user interfaces (3DUI)*, 2016.
- [10] L. Yang, J. Huang, T. Feng, W. Hong-An and D. Guo-Zhong, "Gesture interaction in virtual reality," *Virtual Reality & Intelligent Hardware*, vol. 1, no. 1, pp. 84-112, 2019.
- [11] S. Batra and S. Sachdeva, "Organizing standardized electronic healthcare records data for mining," *Health Policy and Technology*, vol. 5, no. 3, pp. 226-242, 2016.
- [12] M. R. Mine, "Virtual environment interaction techniques," *UNC Chapel Hill computer science technical report TR950-18*, pp. 507248-2, 1995.
- [13] M. Nabiyouni, B. Laha and D. A. Bowman, "Poster: Designing effective travel techniques with bare-hand interaction," in *2014 IEEE Symposium on 3D User Interfaces (3DUI)*, 2014.
- [14] T. Ni, D. A. Bowman, C. North and R. P. McMahan, "Design and evaluation of freehand menu selection interfaces using tilt and pinch gestures," *International Journal of Human-Computer Studies*, vol. 69, no. 9, pp. 551-562, 2011.
- [15] A. Kulshreshtha and J. J. LaViola Jr, "Exploring the usefulness of finger-based 3D gesture menu selection," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2014.
- [16] N. Beattie, B. Horan and S. McKenzie, "Taking the LEAP with the Oculus HMD and CAD-Plucking at thin Air?," *Procedia Technology*, vol. 20, pp. 149-154, 2015.
- [17] P. W. Lee, H. Y. Wang, Y. C. Tung, J. W. Lin and A. Valstar, "TranSection: hand-based interaction for playing a game within a virtual reality game," in *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*, 2015.
- [18] P. Salomoni, C. Prandi, M. Rocetti, L. Casanova, L. Marchetti and G. Marfia, "Diegetic user interfaces for virtual environments with HMDs: a user experience study with oculus rift," *Journal on Multimodal User Interfaces*, vol. 11, no. 2, pp. 173-184, 2017.
- [19] R. Codd-Downey and W. Stuerzlinger, "LeapLook: a free-hand gestural travel technique using the Leap Motion finger tracker," in *Proceedings of the 2nd ACM symposium on Spatial user interaction*, 2014.
- [20] C. Khundam, "First person movement control with palm normal and hand gesture interaction in virtual reality," in *12th International Joint Conference on Computer Science and Software Engineering (JCSSE)*, 2015.
- [21] Y. Sato, M. Saito and H. Koike, "Real-time input of 3D pose and gestures of a user's hand and its applications for HCI," in *Proceedings IEEE Virtual Reality 2001*, 2001.
- [22] J. Chastine, N. Kosoris and J. Skelton, "A study of gesture-based first person control," in *Proceedings of CGAMES'2013 USA*, 2013.
- [23] S. Greuter and D. J. Roberts, "Spacewalk: Movement and interaction in virtual space with commodity hardware," in *Proceedings of the 2014 Conference on Interactive Entertainment*, 2014.
- [24] S. Weibel, U. Bockholt and J. Keil, "Design criteria for AR-based training of maintenance and assembly tasks," in *International Conference on Virtual and Mixed Reality*, 2011.
- [25] J. Araullo and L. E. Potter, "Experiences using emerging technology," in *Proceedings of the 26th Australian Computer-Human Interaction Conference on Designing Futures: The Future of Design*, 2014.
- [26] B. Duy Khuat, D. Thai Phung, H. Thi Thu Pham, A. Ngoc Bui and S. Tung Ngo, "Vietnamese sign language detection using Mediapipe," in *10th International Conference on Software and Computer Applications*, 2021.
- [27] S. Batra, H. Sharma, W. Boulila, V. Arya, P. Srivastava, M. Z. Khan and M. Krichen, "An Intelligent Sensor Based Decision Support System for Diagnosing Pulmonary Ailment through Standardized Chest X-ray Scans," *Sensors*, vol. 22, no. 19, p. 7474, 2022.
- [28] A. Pathak, S. Batra and V. Sharma, "An Assessment of the Missing Data Imputation Techniques for COVID-19 Data," in *Proceedings of 3rd International Conference on Machine Learning, Advances in Computing, Renewable Energy and Communication: MARC 2021*, 2022.
- [29] J. Koch, M. Gomse and T. Schüppstuhl, "Digital game-based examination for sensor placement in context of an Industry 4.0 lecture using the Unity 3D engine—a case study," *Procedia Manufacturing*, vol. 55, pp. 563-570, 2021.
- [30] C. Boletsis and J. E. Cedergren, "VR locomotion in the new era of virtual reality: an empirical comparison of prevalent techniques," *Advances in Human-Computer Interaction*, 2019.
- [31] W. A. IJsselsteijn, Y. A. De Kort and K. Poels, "The game experience questionnaire," 2013.
- [32] G. A. Lee, A. Dünser, S. Kim and M. Billinghurst, "CityViewAR: A mobile outdoor AR application for city visualization," in *2012 IEEE international symposium on mixed and augmented reality-arts, media, and humanities (ISMAR-AMH)*, 2012.
- [33] G. A. Lee, A. Dünser, A. Nassani and M. Billinghurst, "AntarcticAR: An outdoor AR experience of a virtual tour to Antarctica," in *2013 IEEE International Symposium on Mixed and Augmented Reality-Arts, Media, and Humanities (ISMAR-AMH)*, 2013.
- [34] P. Apostolellis and D. A. Bowman, "Evaluating the effects of orchestrated, game-based learning in virtual environments for informal education," in *Proceedings of the 11th Conference on Advances in Computer Entertainment Technology*, 2014.
- [35] I. Ahmed, V. Harjunen, G. Jacucci, E. Hoggan, N. Ravaja and M. M. Spapé, "Reach out and touch me: Effects of four distinct haptic technologies on affective touch in virtual reality," in *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, 2016.
- [36] F. Meijer, B. L. Geudeke and E. L. Van den Broek, "Navigating through virtual environments: Visual realism improves spatial cognition," *CyberPsychology & Behavior*, vol. 12, no. 5, pp. 517-521, 2009.
- [37] M. Nabiyouni and D. A. Bowman, "An evaluation of the effects of hyper-natural components of interaction fidelity on locomotion performance in virtual reality," in *Proceedings of the 25th International Conference on Artificial Reality and Telexistence and 20th Eurographics Symposium on Virtual Environments*, 2015.
- [38] J. Schild, J. LaViola and M. Masuch, "Understanding user experience in stereoscopic 3D games," in *Proceedings of the SIGCHI Conference on human factors in computing systems*, 2012.

# Implementation of Big Data Privacy Preservation Technique for Electronic Health Records in Multivendor Environment

Ganesh Dagadu Puri, D. Haritha

Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation,  
Vaddeswaram, AP, India

**Abstract**—Various diagnostic health data formats and standards include both structured and unstructured data. Sensitive information contained in such metadata requires the development of specific approaches that can combine methods and techniques that can extract and reconcile the information hidden in such data. However, when this data needs to be processed and used for other reasons, there are still many obstacles and concerns to overcome. Modern approaches based on machine learning including big data analytics, assist in the information refinement process for later use of clinical evidence. These strategies consist of transforming various data into standard formats in specific scenarios. In fact, in order to conform to these rules, only de-identified diagnostic and personal data may be handled for secondary analysis, especially when information is distributed or transferred across institutions. This paper proposes big data privacy preservation techniques using various privacy functions. This research focused on secure data distribution as well as security access control to revoke the malicious activity or similarity attacks from end-user. The various privacy preservation techniques such as data anonymization, generalization, random permutation, k-anonymity, bucketization, l-diversity with slicing approach have been proposed during the data distribution. The efficiency of system has been evaluated in Hadoop distributed file system (HDFS) with numerous experiments. The results obtained from different experiments show that the computation should be changed when changing k-anonymity and l-diversity. As a result, the proposed system offers greater efficiency in Hadoop environments by reducing execution time by 15% to 18% and provides a higher level of access control security than other security algorithms.

**Keywords**—Privacy preservation; data privacy; data distribution; anonymization; slicing; privacy attacks; HDFS

## I. INTRODUCTION

The widespread access to virtual health documents and assets controlled through Health Information Systems (HISs) enables to strengthen scientific studies, support care centers through medical education. It also assists various groups for fitness methods through management and governance (e.g., scientific audits for high-satisfactory development and care coordination) [1], [2]. Furthermore, health data is often processed and linked with different data sources, along with facts from scientific trials, having allowed for the harvesting of extra insights which might be beneficial. It is now no longer most effective for scientific practices, however additionally for

designing and improving facts fitness structures and contributing to greater efficient policymaking [3]. While the benefits of secondary use of studies and nursing data makes vital for boosting the high-satisfactory of treatment. There are nevertheless several uncertainties regarding how this data is accessible, through whom, and under what conditions. Data sharing (in particular while it includes private and sensitive attributes and is made accessible to the third party enterprise or geographical region in which it's miles engendered) can bring about a lack of privacy for individuals, along with users and health professionals, further to the requirement to gain earlier than taking part in research [4]. As a result, data safety and affected person privacy are vital demanding situations to address.

Another problem in processing health data for secondary applications is informational discrepancies and diversity in clinical records and data. This is related to the lack of uniform data representation. Hospital Information System (HIS) handles both structured and unstructured documents. Although there are still many unresolved issues that may hinder the use of hybrid materials, especially disjointed health data, they are made possible by leveraging and efficiently integrating structured and semi-structured health information. There is a distinct position arising from their complementary use. In a single, structured way while considering data responsiveness, privacy can hamper, and ethical concerns raise [5]. Digital evidence should be transformed into a unified, standardized, and codified representation using modern methods such as big data analysis [6]. Proper use of this data requires the use of appropriate anonymization procedures. This paper presents a novel integrated architecture for collecting clinical data from heterogeneous sources and transforming them into formats useful for clinical secondary use while adhering to the above requirements. Proper distribution of attribute weightage is also important [7].

The main contributions of this paper are:

- Protect the distribution of large heterogeneous datasets with a novel approach to maintaining privacy in Hadoop environments.
- Defeat various network and database attacks such as SQL injection, collusion attacks and similarity attacks.

The rest of Section II describes state-of-the-art systems that demonstrate previous work by various authors. The Section III

describes the research methods used for the proposed system, including a detailed description of the proposed architecture and execution flow. Section IV provides a description of the algorithm and determines how to handle large data distribution tasks. This section describes data protection against internal attacks and data security algorithms against external attacks. Section V focuses on the results and detailed discussion. Extensive experimental analysis and obtained results are defined in tabular and graphical formats. Finally, Section VI describes the conclusions of the proposed system and future issues.

## II. LITERATURE SURVEY

This section describes various state-of-the-art systems used by previous researchers. According to [8], the newly proposed Secured Map Reduce (SMR) layer introduces a security and privacy layer between HDFS and the MR (Map Reduce) layer, and this approach is known as SMR model. A major value in this work is to facilitate data exchange for knowledge mining. This architecture ensures privacy and security for data consumers, addresses privacy scaling issues, and maintains a trade-off between privacy and utility. SMR models significantly reduce runtime and information loss compared to traditional techniques, and minimizes CPU and memory consumption. According to the work proposed in [9][10], current PPDM strategies have been exhaustively investigated and classified based on data modification methods. This is the researcher's main contribution and will help researchers in the field to fully understand the PPDM. In addition, they compared and considered the advantages and limitations of various PPDM approaches. The vast increase in customer data retention has spawned a new field of research known as privacy-preserving data mining (PPDM). The fundamental challenge of PPDM is to modify data using specific techniques and create powerful data mining models of the modified data while meeting specified privacy requirements and ensuring that information is available for intended data analysis activities. Current review studies aim to leverage data mining jobs without compromising the security of people's sensitive information, especially at the record level.

In the research work proposed in [11], the authors provide a well-designed taxonomy that allows systematic and rigorous classification of this difficult research subject. Recently, the term "big data" has become popular. The proliferation of social networks, the Internet of Things (IoT), and the outsourcing of cloud computing have created an incredible amount, velocity, and variety of data. According to [12], authors proposed Mondrian-based k-anonymity method. A deep neural network (DNN)-based architecture is presented to protect the privacy of high-dimensional data. Experimental results show that the proposed method reduces data information loss while preserving privacy. Many companies actively or passively collect data from consumers. It also collects personal data from various databases.

This data includes personally identifiable information (PII) that can be used to identify an individual. Data analysts and researchers have paid much attention to protecting privacy in the explosion of data for big data and cloud computing.

Numerous data anonymization strategies and DNN privacy models have been thoroughly researched.

In research activities [13], data analysts and academics have paid much attention to privacy-aware data distribution for big data and cloud computing. Various data anonymization approaches and DNN models have been thoroughly researched to protect privacy. A public identity-based PDP protocol for secure data storage helps to protect the privacy of many users. This approach allows TPA to correctly assess the integrity of group-shared data. According to [14], it is a privacy-preserving cloud-based mobile multimedia data exchange system with attribute names and values for each attribute, and only attribute names are visible in access policies. However, the attribute values are included in the cipher text. Encryption has two phases online and offline. Data owners can prepare the intermediate cipher text components in an offline step. After receiving a specific access policy and multimedia data encryption request, the data owner can quickly create the final legal cryptogram in an online phase. Most of the processing costs for verification testing and decryption are offloaded to cloud servers using a decryption outsourcing approach. According to the safety case, PPCMM is adaptively safe in the standard model.

In research work [15], three strategies are used to ensure data confidentiality and integrity. It describes homomorphic encryption, order-preserving encryption schemes, and attribute-based encryption. These strategies are best used in the cloud to ensure privacy. It is also ideal for big data to maintain efficiency and scalability for huge datasets for decision making. Big data is a vast accumulation of enormous data sets that cannot be analyzed by ordinary computing techniques. Big data is therefore a vast amount of rapidly changing data with mixed data types. According to [16], IoT deployments in many industries meet the privacy challenges faced by IoT in resource-constrained devices. It provides an opportunity to address some of the uses of blockchain in various fields and IoT privacy concerns. Based on the utilization of blockchain in IoT, authors proposed various research studies. This study aims to review current research on blockchain applications in IoT for privacy protection. After reviewing current solutions, it was determined that blockchain is the most effective way to avoid identity disclosure, surveillance, and tracking in IoT.

According to [17], a new privacy utility approach uses lightweight elliptic curve cryptography (ECC) to protect privacy and particle swarm optimization (PSO) clustering to maintain utility. PSO is used to cluster datasets and ECC is used to ensure confidentiality of clustered datasets. The proposed method is tested on medical datasets and compared to other methods based on various performance criteria such as clustering accuracy, F-measure, data usefulness, and privacy metrics. According to [18], the various impacts of privacy laws on forensic investigations of embedded devices, the role of anti-forensics, and proposals for embedded forensic investigation guidelines and initiatives to address privacy concerns.

They also proposed a SMART system that enables built-in digital forensic investigations to protect privacy at every level of traditional forensic frameworks. This protects cooperation

with unincorporated owners of embedded computers in cybercrime investigations. According to Slawomir Goryczka et al. [19], it is considered an insider attack against a number of data providers that insert records and attempts to take closure through covert attacks against records inserted by other providers. In this work, the authors proposed a secure multiparty computing protocol for ensuring privacy across multiple data providers. A research paper proposed in [20] identifies privacy issues related to big data and its use. It has been suggested that processing different types of data through different channels poses different threats to user privacy. The authors of [21] emphasized a no-delay framework for the release of medical records to protect privacy. The usefulness of published data is enhanced by a late validation approach. Similarity and skewness attacks could be possible when forming sensitive value groups for publishing records. In a research paper [22], [23], the authors proposed a framework for processing streaming data and measuring term similarity within groups using standard means and addressed the similarity attack issue.

### III. PROPOSED SYSTEM DESIGN

Fig. 1 below shows the multivendor environment and different constraints for data collection. First, it collects synthetic and real-time data from a variety of sources, including medical systems, historical data, and runtime data collection from various web applications. This data may include sensitive user information. Also, records collected; need to create a privacy view when sending data. Malicious user can predict the data of provider by using background knowledge. There is need of the study that defines some security policies on the pre-trained model for secure data distribution. The problem of database attacks and privacy violations on privacy view works like a data hiding technique should be eliminated.

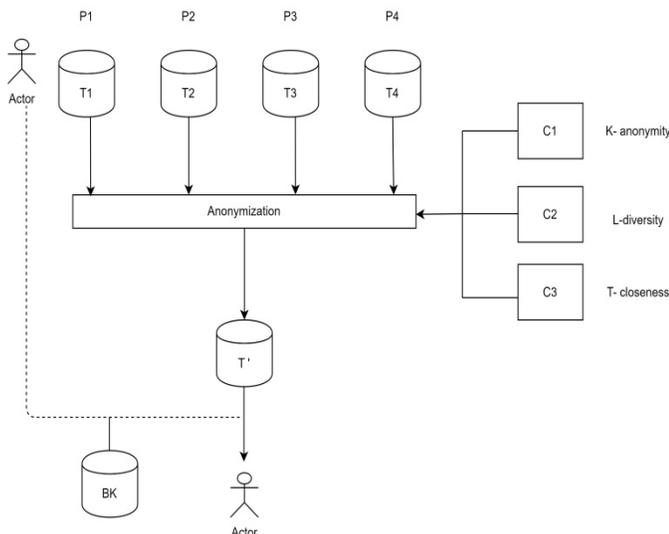


Fig. 1. Multivendor environment and different constraints for data collection

Fig. 2 shows the proposed system architecture with strategic execution. In data providers block, data collected from a variety of sources. It includes various systems such as medical systems, historical data, and runtime data collection

from various web applications. As data collected is very large and considered as big data, nodes are added to process such large data. Using HDFS name nodes are added for processing of data as shown in Fig. 2. Data nodes will provide information for processing to generate privacy view and at the same time detect attack and attacker. Name nodes are keeping directory view of all files in HDFS. Data nodes are sending information to these name nodes and respond to name node in all file operation of privacy view and attack prevention system. This data may include sensitive user information. This study defines some security policies on the pre-trained model for secure data distribution. Privacy view works like a data hiding technique that eliminates the problem of database attacks and privacy violations.

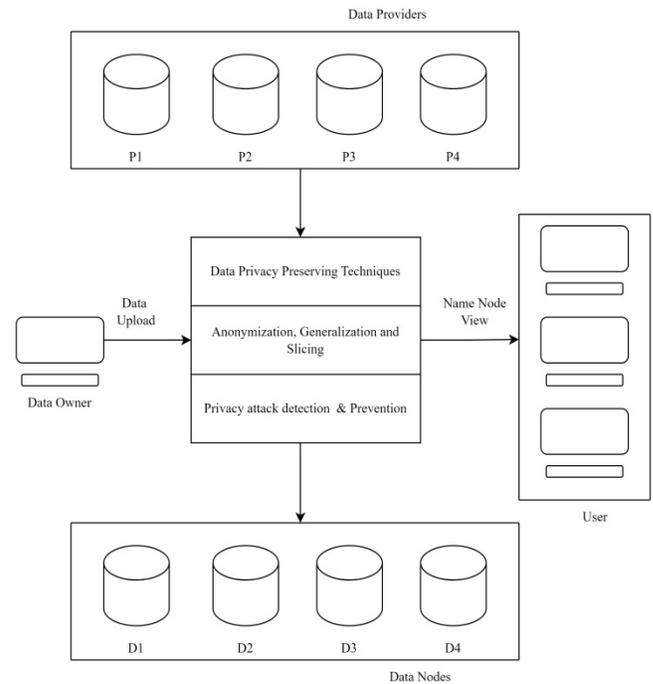


Fig. 2. System architecture for privacy preservation and privacy attack detection in distributed environment

#### A. Techniques for Maintaining Privacy

Anonymization, generalization and slicing are the main methods used in the proposed model. In the anonymization, L-diversity and K-anonymity methods are used. To increase the anonymization level, top-down and bottom-up generalization is used. Slicing basically relies on splitting attributes and tuples. For attribute split (vertical split), split data as {name},{age-zip},{gender} and for tuple split (horizontal split), split as {t1,t2,t3,t4,t5,t6}. In attribute partitioning, age and zip code are quasi-identifiers (QIs), so they are highly correlated and thus partitioned together. These QIs may be known to the attacker. A tuple partitioning system should check the sensitive attribute (SA) column for L-diversity. Equivalence groups are created using K values of quasi identifiers in such a way that they are indistinguishable from each other. In that equivalence class L different sensitive values will be added to make it impossible to identify value of individual from anonymized data.

## B. Privacy Attack Detection and Prevention

Another important feature of this system is the detection of privacy attacks and the prevention of the use of defined algorithms. End users can perform insider, collusion, sql injection and similarity attacks by making minor changes to update the actual value and making it available to another user. Using privacy protection and fingerprint generation technology, such attacks can be easily detected and prevented efficiently

## IV. ALGORITHM DESIGN

### A. Privacy View Generation

It is similar to one-way hash functions to generate the privacy view. The goal of the algorithm presented is to keep sensitive data secure and avoid privacy intrusions. As a result, anonymous views on miniature buckets are generated. In algorithm step 1 to step 6 are used to read the input record wise and apply generalization on quasi-identifiers of the records. While performing anonymization on entire set of quasi-identifiers, validation of that is done using K-anonymity. The records which do not satisfy the criteria of anonymization are added in bucket. Permutation is applied on records so that more records get anonymized. Step 7 to step 12 are applied for the pruning and creating final bucket after anonymization. Bucketization is used to avoid leakage of the records in case of privacy preserved view. It stores records which do not satisfy constraints after pruning and permutation methods.

---

#### Algorithm 1: Algorithm for privacy view generation

---

**Input:** Input dataset DSet, total number of data providers Dp, Constraint policy C {K\_Anonymity, L\_Diversity}

**Output:** Privacy view (NT\*) with selective provider

**Step 1:** foreach (DSet till null)

**Step 2:** foreach (col in table)

    foreach (row in table)

**Step 3:** Select quasi identifier (QiF) and set of sensitive attributes (S\_Att)

**Step 4:** Executes generalization to classify the tuples in QiF with multiple groups

**Step 5:** Perform anonymization on entire set of attributes

**Step 6:** While (validate data privacy(DSet, Dp, C) = 0) do

    if (DSet[i] ← DSet) validated with QiF then

        add D[i] till K-anonymity

    else break;

    Bucket\_List(i1) → DSet;

**Step 7:** Apply permutation on dataset (DSet[i]=I( null-1))

**Step 8:** Apply Pruning on(DSet)

**Step 9:** Execute step 1,2,3 on Bucket\_List (i1)

**Step 10:** if (C != (DSet) && (Dp # 1))

    Bucket(i2) → Bucket\_List (i1(j))

**Step 11:** Show (Bucket\_List (i2)!=null)

**Step 12:** end while

**Step 13:** end for

---

### B. Algorithm 2

The top-down and bottom-up Algorithm 2 is similar to the base-up method. The main difference is in how coalition checks are performed, starting with 0-foe and working up.

---

#### Algorithm 2: Algorithm Top down and Bottom up generalization view

---

**Input:** Input dataset DSet, total number of data providers Dp, Constraint policy C {K\_Anonymity, L\_Diversity}

**Output:** Privacy view (NT\*) with selective provider

**Step 1 :** Read data from dataset from bottom set or top set

$$data[ ] = \sum_{n=1}^m (\text{Row } [n])$$

**Step 2:** Check data count with K-anonymity and L-diversity for each block

**Step 3:** calculate the fitness score F\_Score(DataSet[])

**Step 4 :** if (F\_Score >= Th)

    Generate best generalized view as T\*

**Step 5 :** end loop

**Step 6:** return T\*;

---

When an infringement by any foe is detected (early stop) or all m-policies are examined, the algorithm comes to a cessation. The algorithm represents the basic idea of a bottom-up speculating approach. Using K-anonymity and L-diversity for each block in the dataset, fitness score is checked to generate the best generalized view of privacy preservation. This score is checked against the threshold value for the anonymization.

## V. RESULTS AND DISCUSSIONS

The proposed system is implemented using Java 1.8 and NetBeans 8.0 in an open source Hadoop 2.0 environment. The Intel 2.7 GHz hardware setup is done with 12 GB of RAM. The Hadoop setup is done with a name node and two data nodes using a MapReduce process. According to the problem description, the results obtained are demonstrated using a standalone machine and a Hadoop system. The input dataset size is 101850 instances for both experiments, with and without Hadoop environment. Anonymized views are generated using a definition of C constraints, including K-anonymity and L-diversity. Execution time is measured in milliseconds. Publishing the records using constraints is tedious task if the input data is large. In many cases data need to be preprocessed before giving to the privacy preservation system. This increased data with the need of preprocessing and formatting cannot be executed timely by existing infrastructure and framework.

TABLE I. EXECUTION TIME FOR TRADITIONAL MACHINE WITH VARIOUS L AND K VALUES

Measures	Slicing	Bucket	Final Bucket
L=8 and K=10	11586	1109	184
L=9 and K=12	11776	1316	187
L=10 and K=15	13330	1541	182
L=11 and K=13	11891	1257	188
L=12 and K=15	12950	1550	388

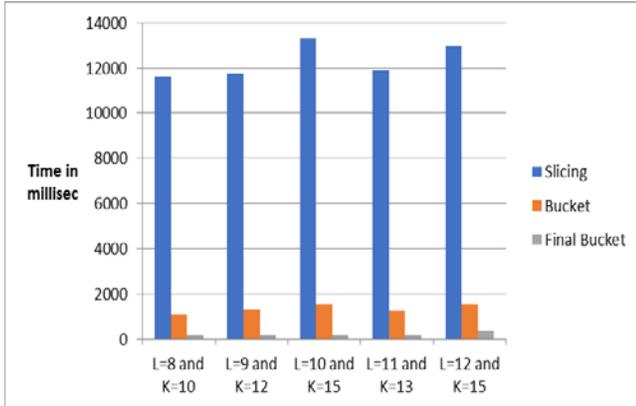


Fig. 3. Execution time for traditional machine with various L and K values

Table 1 and Fig. 3 show the time required to generate T\* views as slicing, bucket generation and final bucket generation. This experiment was conducted on a typical configuration of a single machine. As a result, the execution time of all three processes increased even as the values of L and K increased.

TABLE II. EXECUTION TIME FOR HDFS WITH VARIOUS L AND K VALUES

Measures	Slicing	Bucket	Final Bucket
L=8 and K=10	9827	225	222
L=9 and K=12	7178	262	220
L=10 and K=15	8773	271	186
L=11 and K=13	9354	244	223

After the privacy view generation, few records do not satisfy the constraints set by L-diversity and K-anonymity. If these records are dropped in the system, there is significant loss of information. In this proposed system, these records are stored in the bucket. Bucket is storage area where we can apply the privacy view generation constraints again. Few records may not satisfy the constraints in this stage also. Again dropped records are stored in the final bucket and applied with L-diversity and K-anonymity criteria. For different values of K and L execution time is varying. This execution time is measured in milliseconds.

Fig. 4 shows the time required to generate T\* views using different K and L values mentioned in Table 2. Graph shows the execution time in the Hadoop environment.

Execution time on Hadoop is about 15-18% faster than on a standalone machine with a similar dataset. Increasing the L and K value will increase the execution time subsequently on traditional machine. On HDFS also there is increase in execution time with the increase in these values. Time for privacy view generation, bucket formation and final bucket formation is considered in the execution time.

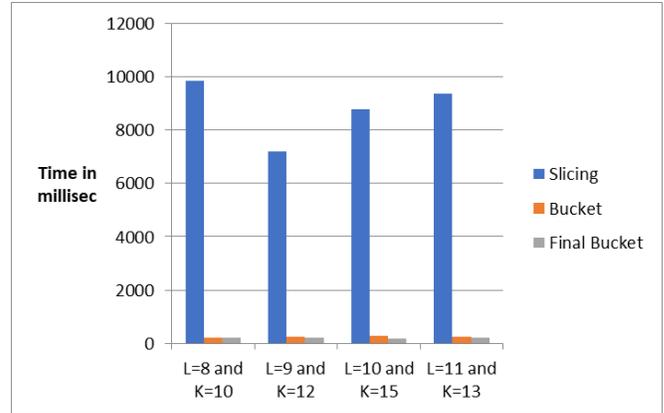


Fig. 4. Execution time for HDFS with various L and K values

TABLE III. EXECUTION TIME FOR TRADITIONAL MACHINE WITH CONSTANT (L=8)

Measures	Slicing	Bucket	Final Bucket
L=8 and K=10	11586	1109	184
L=8 and K=11	12994	299	174
L=8 and K=12	13110	1180	180
L=8 and K=13	11123	1229	196

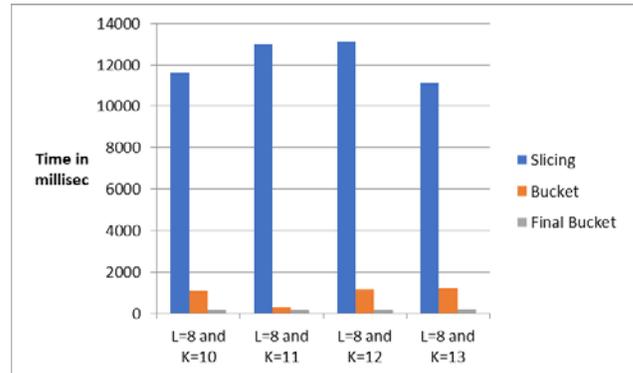


Fig. 5. Execution time for traditional machine with constant (L=8)

Table 3 and Table 4 are enlisting the values for constant L value. Experiment is carried out using constant value 8. In this experiment, value of L-diversity is kept constant and varying k-anonymity values to illustrate the time required generating anonymized views. In this experiment, there is some variation in run time when both values are changed. As only K value is changed, more number of records will appear and available for creation of L-diverse sensitive group. So with increase in k values execution time is reduced.

TABLE IV. EXECUTION TIME FOR HDFS WITH CONSTANT (L=8)

Measures	Slicing	Bucket	Final Bucket
L=8 and K=10	8735	263	208
L=8 and K=11	10362	325	214
L=8 and K=12	9861	281	227
L=8 and K=13	7675	238	213

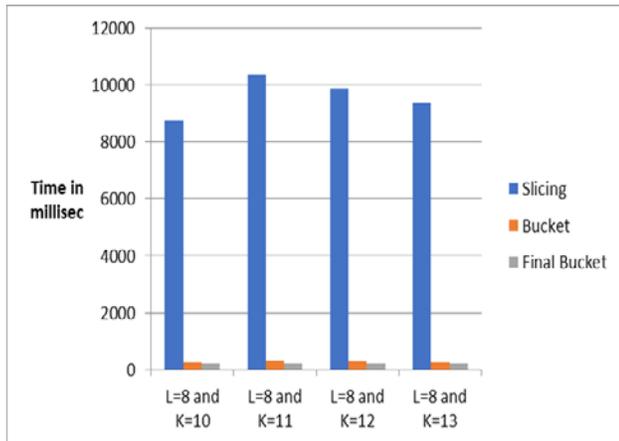


Fig. 6. Execution time for HDFS with constant L-diversity

The Fig. 6 describes the similar execution according to Fig. 5 in Hadoop environment. Almost 24% execution time is reduced using HDFS based a parallel execution.

TABLE V. EXECUTION TIME FOR TRADITIONAL MACHINE WITH CONSTANT (K=15)

Measures	Slicing	Bucket	Final Bucket
L=10 and K=15	13330	1541	182
L=11 and K=15	9500	1534	174
L=12 and K=15	8184	1540	188
L=13 and K=15	10424	1601	192

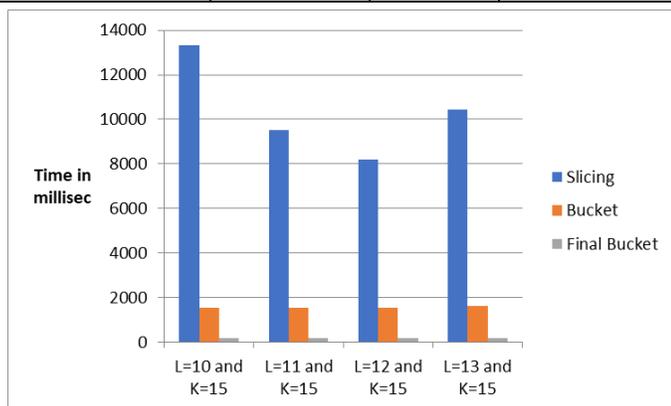


Fig. 7. Execution time for traditional machine with constant k-anonymity

Execution time for traditional machine with constant (K=15) is listed out in Table 5. Fig. 7 shows the generation of T\* views with constant K-anonymity and different L-diversity using a standalone machine. The time required to create the

final anonymized view of slices and buckets is reduced, even if the K and L values change. As multiple providers are providing electronic health records with disease as sensitive attribute, insider attack is possible. M-privacy algorithm takes care of collusion attack which takes place in multivendor environment [19]. But M-privacy algorithm cannot work on large scale data. In this research slicing and bucketization techniques are applied on big data using hadoop distributed file system.

TABLE VI. EXECUTION TIME FOR HDFS WITH CONSTANT K AND DIFFERENT L VALUES

Measures	Slicing	Bucket	Final Bucket
L=10 and K=15	9109	260	461
L=11 and K=15	6555	269	232
L=12 and K=15	7947	272	225
L=13 and K=15	7755	242	219

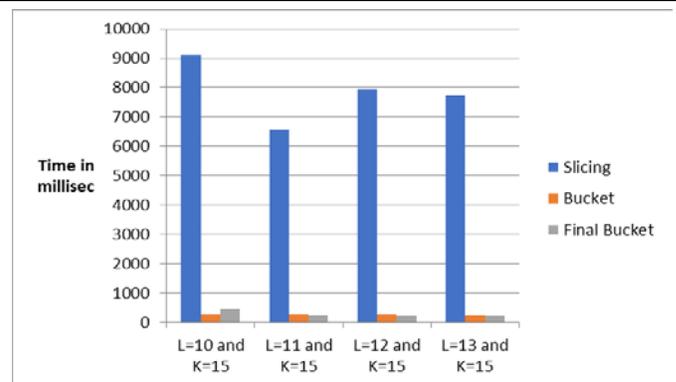


Fig. 8. Execution time for HDFS with constant k and different L values

Table 6 and Fig. 8 above show the generation of privacy views with different values of L-diversity and constant K-anonymity in the Hadoop framework. Slightly changing L-diversity by the k constant does not make much difference. In many cases, the input data cannot satisfy the constraints of C, which can lead to longer times.

## VI. CONCLUSION

In this paper, various privacy-preserving techniques for large-scale health datasets in distributed environments are implemented. Various privacy techniques such as data anonymization, generalization, random permutation, slicing and fingerprinting are used to protect and eliminate privacy attacks. This system provides maximum security in HDFS-based distributed environments and standalone systems. This approach is effective when dealing with real-time data containing sensitive information. Experiments are evaluated on the entire execution using synthetic and real-time healthcare datasets. The system provides 100% privacy with privacy-preserving technology while maintaining the highest accuracy in privacy-based data delivery. Implementation of various machine learning techniques for distributed dynamic data security is a future challenge for the system.

## REFERENCES

- [1] J. Walonoski et al., "Synthesa: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic

- health care record.” *J. Am. Med. Informatics Assoc.*, vol. 25, no. 3, pp. 230–238, 2018, doi: 10.1093/jamia/ocx079.
- [2] M. Asfaw, K. Yitbarek, and J. Gustav, “Emnet : a system for privacy-preserving statistical computing on distributed health data,” *Ep.Liu.Se*, no. June 2015, 2015, [Online]. Available: <http://www.ep.liu.se/ecp/115/006/ecp15115006.pdf>.
- [3] E. Hutchings, M. Loomes, P. Butow, and F. M. Boyle, “A systematic literature review of attitudes towards secondary use and sharing of health administrative and clinical trial data: a focus on consent,” *Syst. Rev.*, vol. 10, no. 1, 2021, doi: 10.1186/s13643-021-01663-z.
- [4] F. Earls and S. Cook, “INTEGRATED ADDENDUM TO ICH E6(R1): GUIDELINE FOR GOOD CLINICAL PRACTICE,” *Child Psychiatry Hum. Dev.*, vol. 13, no. 4, 1983.
- [5] M. Tayefi et al., “Challenges and opportunities beyond structured data in analysis of electronic health records,” *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 13, no. 6. 2021, doi: 10.1002/wics.1549.
- [6] S. Silvestri, A. Esposito, F. Gargiulo, M. Sicuranza, M. Ciampi, and G. De Pietro, “A big data architecture for the extraction and analysis of EHR data,” 2019, doi: 10.1109/SERVICES.2019.00082.
- [7] G. D. Puri and D. Haritha, “Improving Privacy Preservation Approach for Healthcare Data using Frequency Distribution of Delicate Information,” vol. 13, no. 9, pp. 82–90, 2022. (DOI) : 10.14569/IJACSA.2022.0130910
- [8] P. Jain, M. Gyanchandani, and N. Khare, “Enhanced Secured Map Reduce layer for Big Data privacy and security,” *J. Big Data*, vol. 6, no. 1, 2019, doi: 10.1186/s40537-019-0193-4.
- [9] M. Binjubeir, A. A. Ahmed, M. A. Bin Ismail, A. S. Sadiq, and M. Khurram Khan, “Comprehensive survey on big data privacy protection,” *IEEE Access*, vol. 8, 2020, doi: 10.1109/ACCESS.2019.2962368.
- [10] P. GaneshD, P. Dinesh D, and W. Manoj A., “RAID 5 Installation on Linux and Creating File System,” *Int. J. Comput. Appl.*, vol. 85, no. 5, pp. 43–46, 2014, doi: 10.5120/14841-3107.
- [11] H. Y. Tran and J. Hu, “Privacy-preserving big data analytics a comprehensive survey,” *J. Parallel Distrib. Comput.*, vol. 134, 2019, doi: 10.1016/j.jpdc.2019.08.007.
- [12] J. Andrew, J. Karthikeyan, and J. Jebastin, “Privacy Preserving Big Data Publication on Cloud Using Mondrian Anonymization Techniques and Deep Neural Networks,” 2019, doi: 10.1109/ICACCS.2019.8728384.
- [13] H. Yan and W. Gui, “Efficient Identity-Based Public Integrity Auditing of Shared Data in Cloud Storage with User Privacy Preserving,” *IEEE Access*, vol. 9, 2021, doi: 10.1109/ACCESS.2021.3066497.
- [14] Q. Li, Y. Tian, Y. Zhang, L. Shen, and J. Guo, “Efficient Privacy-Preserving Access Control of Mobile Multimedia Data in Cloud Computing,” *IEEE Access*, vol. 7, 2019, doi: 10.1109/ACCESS.2019.2939299.
- [15] H. Shekhawat, S. Sharma, and R. Koli, “Privacy-preserving techniques for big data analysis in cloud,” 2019, doi: 10.1109/ICACCP.2019.8882922.
- [16] Z. Iftikhar et al., “Privacy preservation in resource-constrained iot devices using blockchain—a survey,” *Electronics (Switzerland)*, vol. 10, no. 14. 2021, doi: 10.3390/electronics10141732.
- [17] N. Yuvaraj, R. Arshath Raja, and N. V. Kousik, “Privacy Preservation Between Privacy and Utility Using ECC-based PSO Algorithm,” in *Advances in Intelligent Systems and Computing*, 2021, vol. 1172, doi: 10.1007/978-981-15-5566-4\_51.
- [18] J. Pathak, S. Sankaran, and K. Achuthan, “A SMART Goal-based Framework for Privacy Preserving Embedded Forensic Investigations,” 2019, doi: 10.1109/ISED48680.2019.9096232.
- [19] S. Goryczka, L. Xiong, and B. C. M. Fung, “M-Privacy for Collaborative Data Publishing,” *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 10, 2014, doi: 10.1109/TKDE.2013.18.
- [20] G. D. Puri and D. Haritha, “Survey big data analytics, applications and privacy concerns,” *Indian J. Sci. Technol.*, vol. 9, no. 17, 2016, doi: 10.17485/ijst/2016/v9i17/93028.
- [21] S. Kim, M. K. Sung, and Y. D. Chung, “A framework to preserve the privacy of electronic health data streams,” *J. Biomed. Inform.*, vol. 50, pp. 95–106, 2014, doi: 10.1016/j.jbi.2014.03.015.
- [22] G. D. Puri and D. Haritha, “Framework to avoid similarity attack in big streaming data,” *Int. J. Electr. Comput. Eng.*, vol. 8, no. 5, 2018, doi: 10.11591/ijece.v8i5.pp.2920-2925.
- [23] G. D. Puri and D. Haritha, “A novel method for privacy preservation of health data stream,” *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 9, no. 4, pp. 4959–4963, 2020, doi: 10.30534/ijatcse/2020/110942020.

# Churn Customer Estimation Method based on LightGBM for Improving Sales

Kohei Arai<sup>1</sup>, Ikuya Fujikawa<sup>2</sup>, Yusuke Nakagawa<sup>3</sup>, Ryuya Momozaki<sup>4</sup>, Sayuri Ogawa<sup>5</sup>

Information Science Dept., Saga University  
Saga City, Japan<sup>1</sup>  
SIC Co., Ltd, Hakata-ku, Fukuoka City  
Fukuoka, Japan<sup>2, 3, 4, 5</sup>

**Abstract**—Churn customer estimation method is proposed for improving sales. By analyzing the differences between customers who churn and customers who do not churn (returning), we will conduct a customer churn analysis to reduce customer churn and take steps to reduce the number of unique customers. By predicting customers who are likely to defect using decision tree models such as LightGBM, which is a machine learning method, and logistic regression, we will discover important feature values in prediction and utilize the knowledge obtained through Exploratory Data Analysis (EDA). As results for experiments, it is found that the proposed method allows estimation and prediction of churn customers as well as characteristics and behavior of churn customers. Also, it is found that the proposed method is superior to the conventional method, GradientBoostingClassifier (GBC) by around 10%.

**Keywords**—LightGBM (light gradient boosting machine); EDA (exploratory data analysis); churn prediction; linear regression; gradient boosting method; GradientBoostingClassifier: GBC

## I. INTRODUCTION

Churn customer estimation method is very important for improving sales. By analyzing the differences between customers who churn and customers and who do not churn (returning), a customer churn analysis to reduce customer churn is conducted through taking steps to reduce the number of unique customers. By predicting customers who are likely to defect using decision tree models such as LightGBM, which is one of a machine learning method, and logistic regression, for discovering important feature values in prediction and utilize the knowledge obtained through Exploratory Data Analysis (EDA).

In order to predict churn customers, the method based on LightGBM and EDA is proposed here. LightGBM is decision tree gradient boosting frameworks just as of XGBoost method and is convenient and fast machine learning method. Although there are differences in the details of the implementation method, there is no problem in thinking that they are almost the same framework in general. LightGBM is much faster than XGBoost method because it handles continuous values as histograms. XGBoost did not originally have this implementation, but now it is also possible to adopt a histogram-based algorithm with the parameter `tree_method = hist`.

The comparison between XGBoost and LightGBM is also a research topic because gradient boosting is highly practical.

There is "Benchmarking and Optimization of Gradient Boosting Decision Tree Algorithms" published in September 2018 [1]. After testing XGBoost, LightGBM, and Catboost<sup>1</sup>, it is concluded that no method is clearly superior in all situations.

The specific features and advantages of XGBoost and LightGBM is as follows,

- No need to impute missing values
- There is no problem even if there are redundant feature values (even if there are explanatory variables with high correlation, they can be used as they are)
- The difference from random forest is that trees are made in series.

On the other hand, approaches to data analysis can be broadly divided into a "hypothesis verification type" that verifies hypotheses with data and an "exploration type (EDA)" that generates hypotheses from data. Methods of data analysis are roughly divided into CDA: Confirmatory Data Analysis and EDA [2]-[6]. CDA is a general term for analytical methods aimed at hypothesis verification, while EDA is an analytical method aimed at obtaining hypotheses and knowledge from large-scale, multi-general data. EDA does not select explanatory variables in advance and performs exploratory analysis by seeking knowledge from a wide range of subjects. When we actually analyze data, we go back and forth between the hypothesis testing type and the search type to find out what we know.

Data analysis requires setting hypotheses to be verified, and there is nothing to be gained from analysis without hypotheses. However, there are times when a hypothesis cannot be obtained. Therefore, in order to create a hypothesis, it is necessary to look at the data from various angles and explore trends. Therefore, an exploratory data analysis is performed.

EDA can help by making sure stakeholders are asking the right questions. EDA helps answer questions about standard deviations, categorical variables, and confidence intervals. Once EDA is complete and insights are obtained, the features can be used for more sophisticated data analysis and modeling, including machine learning.

The cost of acquiring a new customer is higher than the cost of retaining an existing customer, up to five times as

<sup>1</sup> <https://catboost.ai/en/docs/concepts/python-quickstart>

much. Therefore, lowering the churn rate has a large positive impact on profits. Churn prediction is especially important for subscription-based services. By predicting churn, you can estimate CLTV (customer lifetime value)<sup>2</sup> and measure the growth potential of your business [7]-[19]. Also, customer churn is when customers cancel services such as subscriptions, and revenue churn is, for example, the loss of Monthly Recurring Revenue: MRR at the beginning of the month.

Customer profiling method with Big Data based on Binary Decision Tree: BDT and clustering for sales prediction is proposed and tested with POS: Point of Sales data [20]. Furthermore, a modified Prophet+Optuna prediction method for sales estimations is also proposed [21]. In this study, churn customer estimation method is proposed and examined with POS data for improving further sales.

In the next section, some of previous works are introduced. Then the proposed method for customer churn prediction is described followed by the experiment. Then conclusion and some discussions are described.

## II. PREVIOUS WORKS

The 5:25 rule states that if you reduce customer churn by 5%, your profit margin will improve by 25%. From a medium-to long-term strategy perspective, it is important to implement planned measures after fully considering the balance between the customer retention rate, the defection rate, and the acquisition of new customers. Selling products to new customers requires five times the cost of selling products to existing customers (1:5 rule). Reducing the probability of customer defection and increasing sales of existing customers are important for increasing corporate profits.

It is important to maintain sales to reduce the withdrawal rate related to the top 20% of the treatment menu from the Partley's law<sup>3</sup>. A good way to identify the top 20% is to use a point card. With a point card, it is relatively easy to identify whether a customer is a regular customer or not.

If the new customer development cost is 100, the existing customer retention cost will be 17 to 20. The top 20% of customers account for 60-80% of total sales. Furthermore, in the bottom 30%, the degree of contribution to sales is less than 4%. The top 5% of customers with the highest loyalty often purchase related products. Reducing the defection rate (=increasing the rate of continuous purchases) has a large impact.

If the defection rate drops from 30% to 20%, the company's expected total sales now and in the future will increase by 1.5 times. A 10% increase or decrease in the attrition rate leads to a 50% increase or decrease in sales.

$$CAV = \frac{OD \cdot CNS \cdot CS}{1 - CPR} \quad (1)$$

where CAV: Customer Asset Value, OD: Overall Demand, CNS: Customer Number Share, CS: Customer Share, CPR: Continuous Purchasing Rate.

where, the share of the number of customers is the ratio of customers who purchase the company's products among all customers in the relevant market, and the intra-customer share is the ratio of the company's products to all purchases of the product group by one customer. In addition, 1-continuous purchase rate: customer defection rate = the ratio of customers who purchased the company's products to no longer purchase the company's products.

For existing customers, the largest defection (=low repeat rate) occurs from the first purchase (F1, Frequency = 1) to the second purchase (F2, Frequency = 2). Also, if the purchase at F1 is not a regular purchase, the repeat rate from F1 to F2 is often about 20 to 40%. Furthermore, the repeat rate rises from F2 to F3, F3 to F4, etc., and when it exceeds F3, it rises to about 70 to 90%, and stable repeat earnings can be obtained.

Possible reasons for separation are as follows.

- 1) I did not get the results I wanted or could not get them.
- 2) I felt that the price of the treatment was higher than the benefits obtained (e.g., I was dissatisfied with the cost performance).
- 3) I felt dissatisfaction and anxiety about the company's response, not the treatment.

Therefore, customer defection analysis is necessary. It is necessary to calculate the "customer defection rate", the percentage of customers who did not use the service for the second time or more during a certain period of time, from Customer Relation Management (CRM) data, and to analyze the trend of "what kind of customers are defected". In particular, if the customer abandons the service after using it multiple times, it is necessary to take a customer's purchase history, frequency, and questionnaire.

For example, conducting questionnaires using Google Form, etc., and the "Frequently Asked Questions (FAQ)" page posted on the company's website have a great impact on customer satisfaction. , it is possible to avoid the risk of customers feeling dissatisfied and leaving. In addition, customer information in CRM is not just for approaching repeat customers, but it is necessary to collect and analyze data to grasp the tendency of customers who have already left, and to find out the reasons for leaving.

## III. PROPOSED METHOD

First, customer churn is defined and then features of the customer churn are extracted from the customer data derived from the POS: Point of Sales data.

Customer churn prediction is performed by the following method.

- *Theme setting*: Define business problem and goals to be achieved → Define Before → After with monitorable metrics
- *Analytics design*: Define the built model and necessary data → In many cases, data such as transaction history and CRM (customer relationship management) system

<sup>2</sup> <https://www.cccmk.co.jp/columns/hint3>

<sup>3</sup> <https://magazinn55.exblog.jp/5554516/>

- *Dataset generation:* Preparing data and performing EDA, performing necessary preprocessing to create datasets suitable for machine learning algorithms.
- *Predictive model training and testing:* Train a churn prediction (departure prediction) model using various machine learning algorithms for classification problems → test the learned prediction model

After that customer churn is characterized and estimated based on LightGBM. Meantime, ROC (Receiver Operatorating Characteristic) curve evaluation method<sup>4</sup> is applied to the estimated churn ratio followed by feature importance is analyzed.

Some of the countermeasures are proposed for mitigation of customer churn.

#### IV. EXPERIMENT

##### A. Data Used

We used POS customer data from 1 September 2009 to December 31, 2021. The outline of the data is as follows:

- 1) Total number of customers (persons) 878,181 Number of unique customer IDs
- 2) Total number of cases (cases) 8,857,257 Number of sales item IDs (cut and color are counted as 2 cases, discounts are also counted as 1 case)
- 3) Cancellation of sales (number of cases) 350,017 Number of sales cancellation details

##### B. Definition of Customer Churn

A customer who visited the store in the previous three months did not return to the store in the next three months, and a customer who did not visit the store was defined as a churn. To give an easy-to-understand example, it was defined as "out of the customers who visited the store between January and March, the customers who visited between April and June returned, and the customers who did not visit the store were rejected."

The format of the final churn prediction output is as follows. It is a specification that predicts the probability that each customer will defect in the next three months. In other words, the customer ID and the likelihood of churn are represented as paired data as shown in Table I.

TABLE I. FORMAT FOR OF THE FINAL CHURN PREDICTION OUTPUT

Customer_ID:	Chance_of_Customer_Churn
1	5%
2	50%
3	30%

About 65,000 customers visited all stores from January to March 2021, and customers who visited between April and June returned to the store, and those who did not return to the store.

“0” in Fig. 1 represents recurrence and “1” as customer churn. The overall churn rate was about 42%.

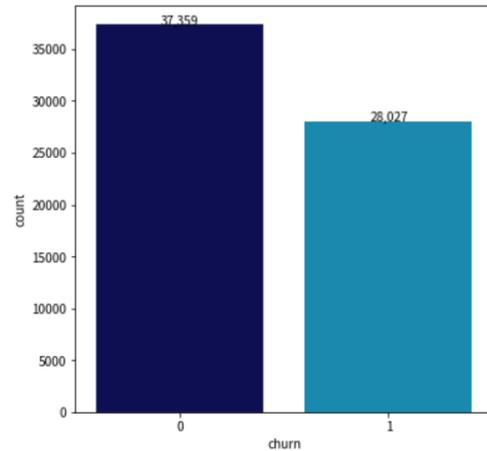


Fig. 1. Overall customer churn

The features used are shown in Table II.

TABLE II. FEATURES USED IN THE POS DATA ANALYSIS

Schema_name	description
customer_id	Customer ID
visits_count	Number of visits
unit_price_ave	Average unit price per store
first_visit_date	Customer's first visit date
last_visit_date	Customer's last visit date
gender	Gender
age	Age(customers_who_do_not_enter_are_0)
distance	Distance_to_the_store_calculated_from_the_zip_code
menu	Categorization_by_menu
unit_price_per_visits_co unt	Average_unit_price_per_visit/number_of_visits

##### C. Preliminary Results

1) *The difference between churn customers and non-churn customers:* The difference between churn customers and non-churn customers was evaluated from the number of visits. The results are shown in Fig. 2. In the figure, orange indicates churn and blue indicates return. The lower the number of visits to the store, the higher the attrition rate, and the higher the number, the lower the attrition rate. There is a marked difference.

<sup>4</sup> <https://zero2one.jp/ai-word/roc-curve-and-auc/>

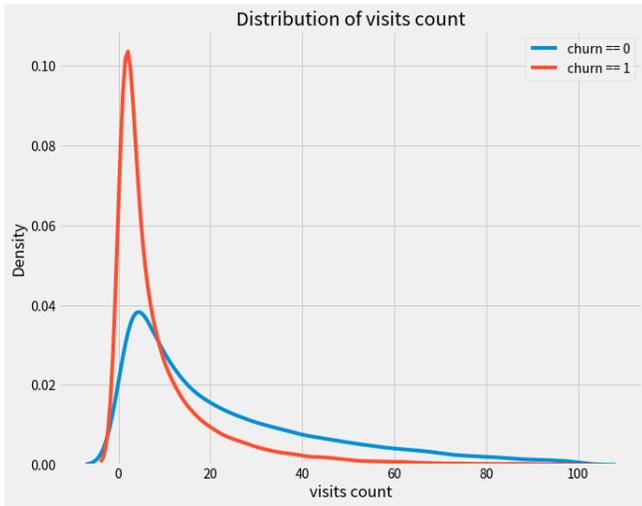


Fig. 2. The difference between churn customers and non-churn customer

2) *Average cost per visit:* Customers with low unit prices have a high churn rate, and customers with high unit prices have a low churn rate. However, there is no big difference depending on the unit price as shown in Fig. 3.



Fig. 3. Average cost per visit dependency against churn rate

3) *Customer's first visit date:* The horizontal axis in Fig. 4 indicates how many days before the first visit to the store from the analysis point. This time, we analyzed customers who visited the store from January to March, so March 31st was the day before. From this, we can see that the churn rate is higher for people who first visited the store recently, and the churn rate is lower for people who first visited the store a long time ago. These differences are significant.

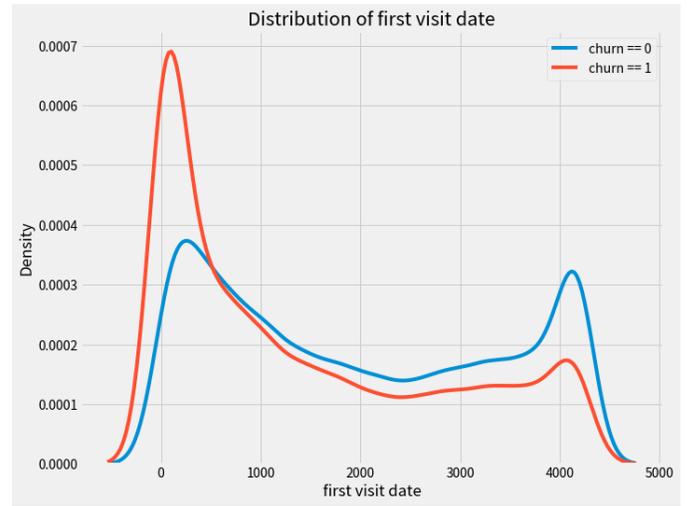


Fig. 4. Customer's first visit date dependency against churn rate

4) *Customer's last visit date:* The horizontal axis in Fig. 5 shows the number of days before the last visit from the point of analysis, just like the date of the first visit. This time, we analyzed customers who visited the store from January to March, so March 31<sup>st</sup> was the day before. From this result, we can see that the withdrawal rate is lower for those who last visited the store more recently, and the withdrawal rate is higher for those who last visited the store more than 50 days ago.

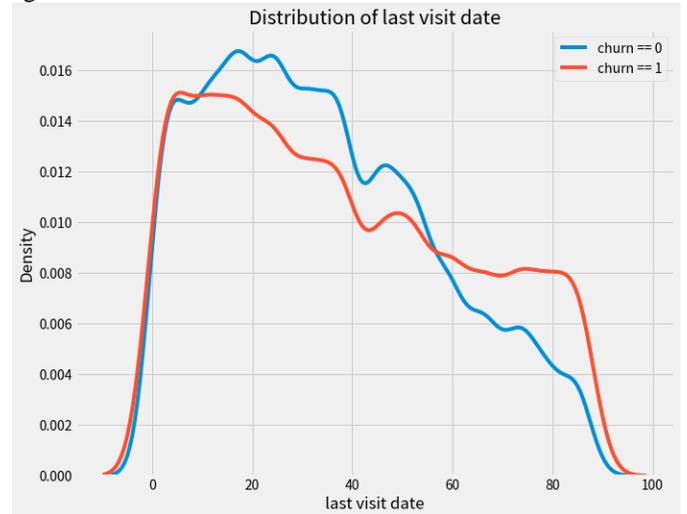


Fig. 5. Customer's last visit date dependency against churn rate

5) *Gender:* The churn rate is lower for men than for women (the churn rate for those entered as women exceeds 60%, but for men it is a little over 50%) as shown in Fig. 6. Customers whose gender is unknown (not entered) have a very low churn rate. The reason for that is unclear.

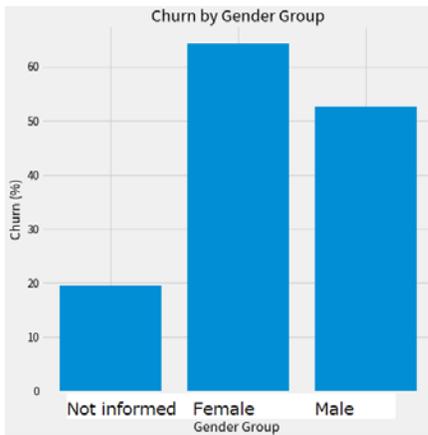


Fig. 6. Gender dependency against churn rate

6) *Age*: The churn rate is high for those in their 20s and 30s and decreases for those in their 50s as shown in Fig. 7.

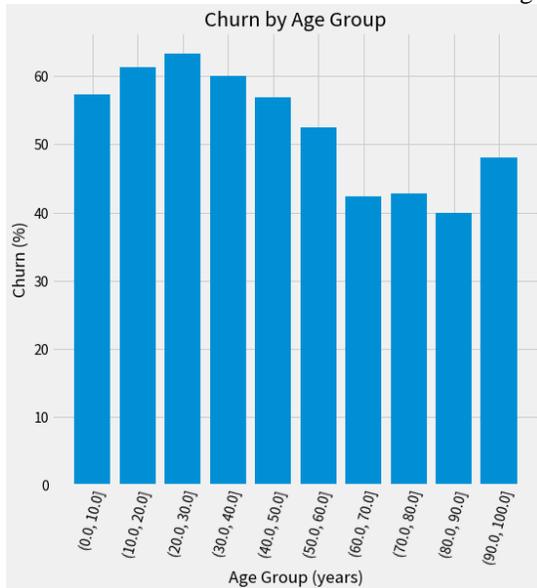


Fig. 7. Age dependency against churn rate

7) *Service menu*: We categorized customers according to the menu they ordered the most and investigated the churn rate. As a result, it was found that the rejection rate for dyeing white hair is very low at around 30%, while the rejection rate for child cuts and school cuts is high as shown in Fig. 8.

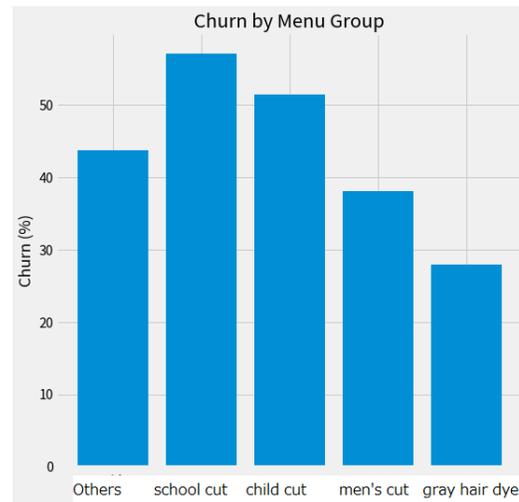


Fig. 8. Service menu dependency against churn rate

8) *Average cost per visit / number of visits*: Fig. 9 shows only those customers whose average unit price per visit/ number of visits is more than 2000 Yen in KDE<sup>5</sup> (Kernel Density Estimation). Customers with this value of 6,000 Yen or more seem to have a slightly higher churn rate. In other words, it seems that the churn rate is high for people who order expensive menus despite the fact that they visit the store less frequently.

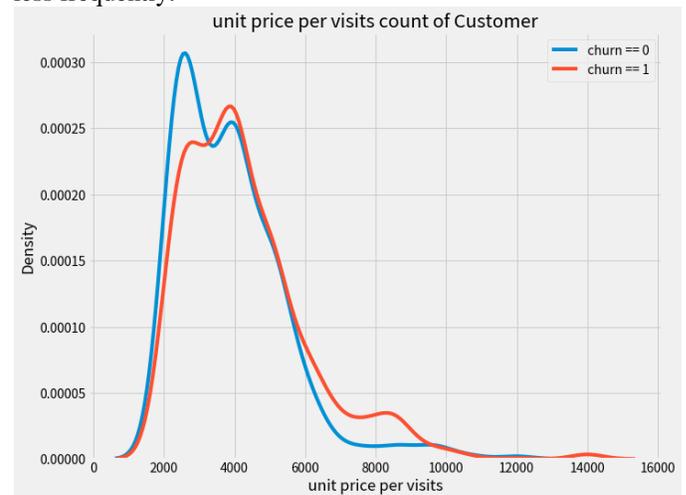


Fig. 9. Characteristics of the average cost per visit / number of visits

<sup>5</sup> <https://cran.r-project.org/web/packages/spNetwork/vignettes/NKDE.html>

D. Customer Churn Prediction

1) *LightGBM based prediction of customer churn*: The results of predicting customer churn using the above feature values (excluding distance to the store) are shown below. Fig. 10 shows the feature value order of customer churn prediction using LightGBM. It can be seen that the number of visits to the store on the day of the first visit has a large effect and is greatly affected to the churn.

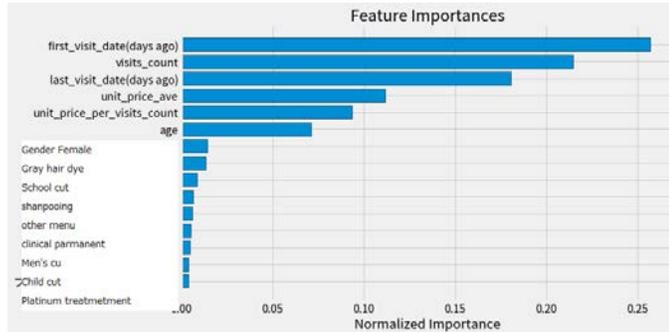


Fig. 10. Feature value order of customer churn prediction using LightGBM

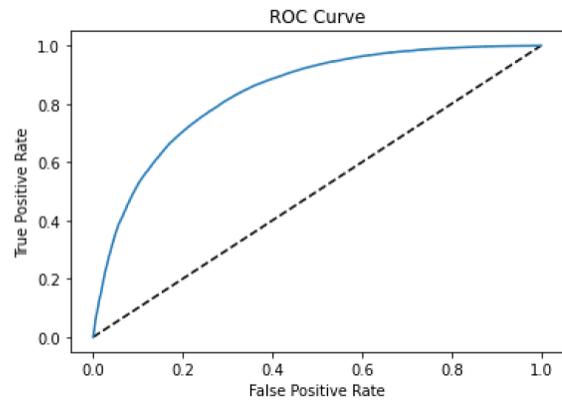
2) *ROC curve evaluation*: ROC curve and Churn pct are evaluated. Each axis represents TPR (True Positive Rate) and FPR (False Positive Rate) and plots the TPR and FPR values when changing the threshold for classifying into Positive and Negative. As shown in Fig. 11, ROC curve and Churn pct (histogram) are seemed reasonable (not perfectly satisfied but marginal). Also, AUC (Area Under the Curve) and logarithmic function of loss are evaluated. As shown in Table III, both show reasonably satisfied values.

TABLE III. AUC AND LOGARITHMIC FUNCTION OF LOSS

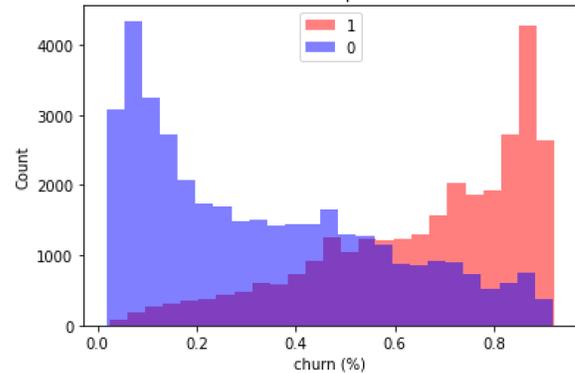
AUC	0.837
Log_loss	0.496

3) *Customer churn characteristics*: Customer churn characteristics are summarized as follows:

- a) *Age*: Younger customers have a higher attrition rate, and those in their 60s to 80s have a lower attrition rate.
- b) *Gender*: Female divorce rate is high.
- c) *Number of visits*: The lower the number, the higher the churn rate.
- d) *Unit price (1 store visit)*: The lower the rate, the higher the withdrawal rate.
- e) *Date of first visit*: Most recent customers (customers who have recently visited for the first time) have a high churn rate.
- f) *Date of last visit*: The withdrawal rate is lower for the most recent visit.
- g) *Menu*: Low withdrawal rate for gray hair dyeing, high withdrawal rate for school and child cuts.
- h) *Distance to stores*: This hardly contributes to the churn rate and seems to depend on the availability of parking lots.



(a) ROC curve



(b) Churn pct (histogram)

Fig. 11. ROC curve and churn pct of the churn rate prediction based on LightGBM

V. CONCLUSION

Churn customer estimation method is proposed for improving sales. By analyzing the differences between customers who churn and customers who do not churn (returning), we conduct a customer churn analysis to reduce customer churn and take steps to reduce the number of unique customers. By predicting customers who are likely to defect using decision tree models such as LightGBM, which is a machine learning method, and logistic regression, we discover important feature values in prediction and utilize the knowledge obtained through EDA.

As results for experiments, it is found that the proposed method allows estimation and prediction of churn customers as well as characteristics and behavior of churn customers. Also, it is found that the proposed method is superior to the conventional method, GradientBoostingClassifier: GBC by around 10%.

FUTURE RESEARCH WORKS

Further investigations are required for improvement of prediction accuracy. We could be able to take measures such as sending DMs and coupons to customers with a 90% chance of churn. In order to increase the accuracy of churn prediction, not only LightGBM but also ensemble models such as Random Forest and logistic regression will be learned, and the accuracy will increase a little more. In addition, this time, we had the customers of all stores who visited the store during a specific

period learn, but if we try to learn for each store without narrowing down the period, a different result may appear.

#### ACKNOWLEDGMENT

The authors would like to thank to Professor Dr. Hiroshi Okumura and Professor Dr. Osamu Fukuda for their valuable discussions.

#### REFERENCES

- [1] Andreea Anghel, Nikolaos Papandreou, Thomas Parnell, Alessandro De Palma, Haralampos Pozidis, Benchmarking and Optimization of Gradient Boosting Decision Tree Algorithms, Workshop on Systems for ML and Open Source Software at NeurIPS 2018, Montreal, Canada, 2018.
- [2] Hirokazu Iwasawa, Yuji Hiramatsu, "EDA (Exploratory Data Analysis) Predictive Modeling with R: For Risk Management Using Machine Learning Tokyo Tosho pp.46-62, 2019.
- [3] Yasuhito Mizoe, "Concept of Exploratory Data Analysis," Estrela, No.65, August 1999, pp.2-8, 1999.
- [4] Mosteller, F. and J.W. Tukey, "Data Analysis and Regression", Addison- Wesley, 1977.
- [5] Noora Kanerva, Jukka Kontto, Maijaliisa Erkkola, Jaakko Nevalainen, Satu Männistö, "Suitability of random forest analysis for epidemiological research: Exploring sociodemographic and lifestyle-related risk factors of overweight in a cross-sectional design." Scandinavian Journal of Public Health, Vol 46(5) pp.557-564, 2018.
- [6] Tukey, J.W., "Exploratory Data Analysis", Addison-Wesley, 1977.
- [7] Stone, Merlin and Shaw, R., "Database marketing". Aldershot, Gower, 1988.
- [8] Peppers, D., and M. Rogers, Enterprise One to One: Tools for Competing in the Interactive Age. New York: Currency Doubleday, 1997.
- [9] Hanssens, D., and D. Parcheta (forthcoming). "Application of Customer Lifetime Value (CLV) to Fast-Moving Consumer Goods.", 2011.
- [10] Nakamura and Higa, Many studies on CLV ... Each paper has different definitions of customer lifetime value, target industries, business models, conditions for calculation, etc. 2011.
- [11] Nakamura and Higa, "There are cases where COCA (Cost of customer acquisition), which is the cost of acquiring customers, is added 2011.
- [12] Berger, P. D.; Nasr, N. I., "Customer lifetime value: Marketing models and applications". Journal of Interactive Marketing 12 (1): 17-30. doi:10.1002/(SICI)1520-6653(199824)12:1<17::AID-DIR3>3.0.CO;2-K 1988.
- [13] Fripp, G., "Marketing Study Guide" Marketing Study Guide, 2014.
- [14] Adapted from "Customer Profitability and Lifetime Value," HBS Note 503-019, 2014..
- [15] Ryals, L., Managing Customers Profitably. ISBN 978-0-470-06063-6. p.85, 2008.
- [16] Gary Cokins, Performance Management: Integrating Strategy Execution, Methodologies, Risk and Analytics. ISBN 978-0-470-44998-1. p. 177, 2009.
- [17] Fader, Peter S and Hardie, Bruce GS and Lee, Ka Lok, "RFM and CLV: Using iso-value curves for customer base analysis". Journal of marketing research (SAGE Publications Sage CA: Los Angeles, CA) 42 (4): 415-430. doi:10.1509%2Fjmk.2005.42.4.415, 2005.
- [18] Tkachenko, Yegor, "Autonomous CRM control via CLV approximation with deep reinforcement learning in discrete and continuous action space". arXiv preprint arXiv:1504.01840. doi:10.48550/arXiv.1504.01840, 2015.
- [19] V. Kumar, Customer Lifetime Value. ISBN 978-1-60198-156-1. p.6, 2008.
- [20] Kohei Arai, Zhang Ming Ming, Ikuya Fujikawa, Yusuke Nakagawa, Tatsuya Momozaki, Sayuri Ogawa, Customer Profiling Method with Big Data based on BDT and Clustering for Sales Prediction, International Journal of Advanced Computer Science and Applications, 13, 7, 22-28, 2022.
- [21] Kohei Arai, Ikuya Fujikawa, Yusuke Nakagawa, Tatsuya Momozaki, Sayuri Ogawa, Modified Prophet+Optuna Prediction Method for Sales Estimations, International Journal of Advanced Computer Science and Applications, 13, 8, 58-63, 2022.

#### AUTHORS' PROFILE

**Kohei Arai**, He received BS, MS and PhD degrees in 1972, 1974 and 1982, respectively. He was with The Institute for Industrial Science and Technology of the University of Tokyo from April 1974 to December 1978 also was with National Space Development Agency of Japan from January, 1979 to March, 1990. During from 1985 to 1987, he was with Canada Centre for Remote Sensing as a Post-Doctoral Fellow of National Science and Engineering Research Council of Canada. He moved to Saga University as a Professor in Department of Information Science on April 1990. He was a councilor for the Aeronautics and Space related to the Technology Committee of the Ministry of Science and Technology during from 1998 to 2000. He was a councilor of Saga University for 2002 and 2003. He also was an executive councilor for the Remote Sensing Society of Japan for 2003 to 2005. He is a Science Council of Japan Special Member since 2012. He was Adjunct Professor of University of Arizona, USA from 1998 2020. He also is Vice Chairman of the Science Commission "A" of ICSU/COSPAR from 2008 to 2018 then he is now award committee member of ICSU/COSPAR since 2018. He is now adjunct professor of Nishi-kyushu University and Kurume Institute Technology since 2018. He wrote 92 books and published 689 journal papers as well as 513 conference papers. He received 70 of awards including ICSU/COSPAR Vikram Sarabhai Medal in 2016, and Science award of Ministry of Mister of Education of Japan in 2015. He is now Editor-in-Chief of IJACSA and IJISA.

<http://teagis.ip.is.saga-u.ac.jp/index.html>

# Privacy Preservation Modelling for Securing Image Data using Novel Ethereum-based Ecosystem

Chhaya S Dule<sup>1</sup>, Dr. Roopashree H.R<sup>2</sup>

Research Scholar, Dept. of CSE<sup>1</sup>

Assistant Professor, Department of CSE, Dayananda Sagar University Bangalore, India<sup>1</sup>

Research Supervisor and Associate Professor<sup>2</sup>

Department of Computer Science and Engineering<sup>2</sup>

GSSS Institute of Engineering and Technology for Women, Mysuru, VTU, Belgaum, Karnataka, India<sup>1,2</sup>

**Abstract**—The broad usage of images in real-time applications demands a cloud infrastructure due to its advantages. Many use cases are built where the image data is shared, sharing becomes the core function, and the medical domain takes its broad advantage. The cloud is a centralized infrastructure for its all-operation usages; it depends mainly on the trusted third party to handle security concerns. Therefore, the privacy preservation of the image data or any data becomes an issue of concern. The distrusted system advantages are achieved using blockchain technology for image data security and privacy concerns. The traditional approaches of the security and privacy models raise many apprehensions as these are designed on the centralized systems of the data sharing mechanisms. It is also observed that large data files are not wisely handled, which demands building a framework model that takes image data and any other data of any size to ensure a dependable optimal security system. This paper presents a framework model to achieve optimal time complexity for securing the privacy aspects of the image data or any other data that uses space optimal file system using distributed security mechanism for both the storage and sharing of the data. The proposed framework model for optimal time complexity and security uses a duplication algorithm using stakeholder agreement to ensure efficient access control to the resources using the cryptographic approach to the Ethereum ecosystem. The performance metric used in the model evaluation includes the degree of availability and efficiency. On benchmarks, it performs well compared to the traditional cloud-built distributed systems. The quantified outcome of the proposed scheme exhibits a 42.5% of reduction in time for data repositioning, a 41.1% of reduction in time for data retrieval, a 34.8% of reduction in operational cost, a 73.9% of reduction in delay, and a 61% faster algorithm execution time in contrast to conventional blockchain method.

**Keywords**—Blockchain; data security; Ethereum; image data; privacy; security

## I. INTRODUCTION

In the real world, many applications may include the Map image, medical images, or any other real-time images to be stored and shared to meet their respective objectives of map-based driving assistants, diagnostic systems for medical professionals, and smart transport systems [1-3]. These applications extract intrinsic information for the functional operation and computation of the applications, including general and private information. Vulnerability to private

information is considered a loophole to the system's robustness [4-5]. Though the digital systems built for real-time applications boot the connivance of operation by suitable ecosystems, the cumulative data storage raises concerns about the considerable data handling challenges [6]. Application engineers and domain experts for an application built on the data storage and sharing required to analyze the data features of interest to meet the application goals. However, the data generator or uploader stakeholder may not be able to keep track of the data that might be useful to them. Digital data sharing provides ease of data analytics that ensures application efficiency and, as a result, yields an automatic and intelligent decision-making scheme for specific applications [7]. Though there exist many advantages of end-to-end systems for data or image data storing and sharing yet, significant challenges exist that require attention by the researchers, and those significant challenges include: 1) storage aspect of the extensive data in terms of time complexity of retrieval and 2) the vital security concerns. The research statistics reveal that data exist in heterogeneous formats, but the images are large in percentages [8]. The images are multi-dimension data and require larger storage space than other data formats. Its processing takes more and more computing storage to process and analyzes, limiting exploiting its usage for the application goals and efficiency. At the same time, vulnerability due to various system loopholes concerns data security and privacy information leakage. It lacks efficient interoperability requirements due to the inefficient trade-off handler capacity of the security models [9]. Most data possess critical and delicate material; therefore, preserving privacy is essential to maintain the stakeholder losses or damages, either financial or reputational.

It is also essential that only the authenticated data be utilized to perform any analytical process, as the tampered image data or other data minimizes the accuracy and reliability of the results of the decision-making systems that may deceive the correctness of the system. The secure and fast system ensures better data interoperability, so sharing the data securely makes the system more robust and effective. The cloud ecosystem is today's most popular choice for flexible data storage and optimal sharing. However, towards the security of the data, the cloud ecosystem adopts practical cryptography and privacy preservation, and access control followed by the appropriate authentication schemes [10]. Irrespective of the various security schemes available for data security and privacy leakage protection, there are no full-proof adaptive

systems. The assumption lacks the reality that storage and data distribution do not have threats if it is through the cloud. This unreliability is because cloud security models largely depend upon a Trusted Third Party (TTP), and TTP suffers many collusion attacks without suitable non-repudiation schemes [11]. Regrettably, there is no typical confirmation instrument for prevailing systems, and there is no operative countermeasure to punish a malicious process in the cloud ecosystem. The most promising distributed technology platform Ethereum is gaining popularity. Recently, it has been used to build many security models or schemes to provide reliable and robust data-sharing systems using distributed databases [12]. The Ethereum-based platform facilitates distributed technology like blockchain. This open-source distributed database mechanism can be exploited to build effective authentication, access control, secure storage, and sharing schemes for an image or other data [13].

Though there are many significant advantages of the Ethereum-based distributed systems, many traditional approaches based on this technology limit their potential effect due to many of the issues that may include: Though there are many significant advantages of the Ethereum-based distributed systems. Still, many of the traditional approaches based on this technology limit their potential effect due to many of the issues that may include: 1) It lacks non-repudiation in the distributed databases as well as does not handle the large data effectively due to non-scalability factor as these distributed data is accessible to all the participating databases as a ledger. 2) Another associated problem is that it lacks control over the data access by the authenticated stakeholder. 3) And last but not least, to design an optimal system to minimize the time complexity and yet be robust enough to ensure system reliability and availability most securely. Literature reveals the fact that these problems are not handled well. Therefore, this paper deals with the two primary objectives: a framework component that balances the time complexity overhead for data security and privacy preservation of images or any other data over a distributed database storage and sharing ecosystem. This concept's prime agenda is also to balance privacy preservation with reduced time complexity associated with the data-sharing process on the distributed scale over cloud networks. To sum up, the contribution of the paper is as follows:

1) The framework model is deployed on the open-source Ethereum distributed database system over the cloud to validate the agreements among the stakeholders of the file storage and sharing process and authenticated access control mechanism.

2) The framework model exploits a unique type of file system, namely Space Optimal File System (SOFS), in the form of a node-to-node collaborative approach that can store the image or any other data to overcome the limitations of the centralized storing and sharing systems.

3) The framework is flexible to work not only on image data. Instead, it can work on any other data of regular or large files to overcome the limitations of the non-scalability in traditional systems built on distributed systems.

4) The cryptographic key mechanism is used for every chunk of the data so that data is not accessible to the SOFS ledgers

5) Finally, it provides a novelty of handling optimal time complexity and data security to ensure real-time feasible data storage and sharing mechanisms.

The paper is presented in six sections. The literature review is described in Section II, followed by a discussion of identified research gap in Section III. Section IV elaborates on the design and deployment aspects of the proposed framework model, an algorithm discussion is carried out in Section V, the performance analysis is described in Section VI, a discussion of the accomplished outcome is carried out in Section VII, and finally, Section VIII concludes the paper

## II. LITERATURE REVIEW

Jiang et al. [14] choose a blockchain for the data store and search on the Ethereum platform by designing a price model using two distinguished stakeholders, namely the data owner, who is awarded for providing the data. Another stakeholder is the miner, who is granted the search operation. This model minimizes the keyword duplication cost to gain a cost advantage. Thus, Debe et al. [15] offer a distributed scheme using an Ethereum-based blockchain system to handle these issues. The system validation against the attack models shows resilience to it. In the work of Hasan et al. [16], a blockchain-based model deals with images and other data to handle the storage and the sharing contract among the stakeholders. A joint study on scalability and Adhoc usage of accountability is considered a research problem by Podgorelec et al. [17] and proposes a concept of state channel as a service (SaaS) that ensures secure distributed connections for off-site chain issues in the payment system. Another significant work in the healthcare domain is Madine et al. [18], which securely uses a specific file system, cryptography, and blockchain-based authorization architecture to share patient records' consent. In the work of Abou-Nassar et al. [19], an interoperable distributed trust model is proposed using blockchain using C# on Ethereum. A practical approach towards the design of "privacy-preserving permissioned blockchain architecture" has been presented by Lin et al. [20] by modifying the Ethereum and customized cryptographic intrinsic elements. The authors, Kumar et al. [21], highlight that along with health care, another domain like cryptosecurity, distributed data collaboration, and immutability take advantage of blockchain technologies. The work of Ullah et al. [22] highlights that though the current cloud ecosystem-based data store system provides many advantages but still lacks data leakage and risk to private information due to the centralized operations and dependency on the TTP, a single point of failure may collapse the system. In Yan et al. [23], a dynamic data upload process and search verification through a fuzzy keyword are proposed. Using E-blockchain and Rivest-Shamir Algorithm (RSA) ensures fairness between the user and the cloud data store. Xiang et al. [24] have presented a data trading mode using blockchain and machine learning by building a contract among the stakeholder by eliminating the TTP. The E-Blockchain is used to meet this requirement in the model proposed by Debe et al. [25], enables a decentralized agreement to establish trust-based sharing among the IoT devices and the fog, and performs well

compared to the existing trust model, which is centralized for its operation. Yet another recent and significant work by Hasan et al. [26] proposes the data chunk transfer process decentralized using blockchain that ensures privacy and confidentiality using a proxy cryptographic approach using a specific file system. Yang et al. [27] presented a layer-based trust approach using the Hyperledger architecture to handle this problem. It has been validated against the attack, namely distributed denial service (DDoS). However, one interesting fact about Ethereum, blockchain, and Smart Contracts is discussed in the work of Chen et al. [28], where the users encounter the threats of resource abuse. Saini et al. [29] presented a framework using Smart Contracts (SC) and blockchain (BC) for access control to secure healthcare-related data. Debe et al. [30] present a scheme that uses BC and SC for a decentralized Bidding process and a reputation system that cost-effectively ensures security. The use of BC and Ethereum with SC is also found in Kaynak et al. [31]. Table I highlights the summary of the above-related studies on security.

TABLE I. SUMMARY OF RELATED WORK

Approaches	Advantage	Limitation
Blockchain-based [14]-[19]	can secure different forms of data	No benchmarking
Blockchain with Cryptography [33]-[37], [39], [40]	Ensure privacy preservation, better access control, secure sharing	Computationally complex process
Homomorphic encryption [42]	Secure, ubiquitous data	Slow execution time
Traditional Public Key Encryption [32][41][38]	Simpler architecture, the supportability of wide varieties of application	Not resistive to dynamic attackers

Thus, this paper aims to fill this gap by designing a generalized framework model that supports image and non-image data using the Ethereum platform and cryptography for privacy preservation in an optimal time complexity way. The next section outlines problems derived from existing approaches.

### III. IDENTIFIED RESEARCH CHALLENGE

After reviewing the existing schemes, specific research challenges have surfaced, which are as follows:

- A centralized server stores the information in most of the existing blockchain-based schemes. Such storage often invites identity theft, privacy leakage, and other associated security issues.
- Adopting complex security architectures requires proper knowledge to handle them in case of unidentified attacks on data. Mishandling of security features by data owners eventually leads to intrusion. At the same time, they also have to depend on adopting trusted third parties, which are equally vulnerable from a data ownership viewpoint.
- Conventional blockchain-based operations cannot manage large-scale data and very often lead to scalability issues that degrade the performance of repositioning and querying the data from miners.

- The majority of existing studies suffer from loosely coupled ownership of data. Once the blockchain stores the data, the availability and reachability of data are far more to other users using malicious access policies.

Hence, the above mentioned research problems have been identified and are subjected to proposed solutions emphasizing data privacy preservation. Apart from this, the storing the image in conventional blockchain is usually centralized and there is higher possibility of intrusion, whereas Ethereum-based approach are decentralized and its control of validation is carried out by multiple nodes with higher accountability. Hence proposed study considers Ethereum-based ecosystem for this purpose. The following section discusses the adopted research methodology.

### IV. RESEARCH METHODOLOGY

This paper proposes a customized framework model that supports image data security for its privacy preservation in the optimal time complexity by exploiting the Ethereum-based distributed data store system characteristics like blockchain.

The summary of the adopted method is as follows: The overall structure of the implementation is classified into four stages. The *first stage* of development is associated with constructing the Ethereum module, which maintains various associated repositories of data identity, timestamp, Secured Hash Algorithm (SHA), and auditor identity. The *second stage* of development is primarily responsible for request assessment and management of metadata. This module further contributes towards applying asymmetric encryption over the split data in adherence to the bandwidth capacity. The idea is to secure the storage units. The *third stage* of development performs authorization to the data requestor while Ethereum records are validated. Finally, the *fourth stage* of the proposed method introduces the smart agreement process, where the transactions are assessed and duplicate records are identified. The module finally performs a key updating process.

The framework aims to support or work on image data or other small to large-size and multi-dimension data to ensure the system's scalability. The core idea of the proposed scheme is to introduce an authorization mechanism for image data from various perspectives of the application using images; it could be for multi-disciplinary applications. It will also eventually mean that considered image data could easily take the shape of high-dimensional data, progressively increasing its challenge during the computational process. In the proposed scheme, the user is facilitated with the privilege to construct a tailored policy towards accessing their intellectual property (image) using an intelligent agreement system when applied with Ethereum blockchain. Further, a specific administrative use is responsible for uploading the comprehensive image data over the proposed storage network. The complete operation of uploading and accessibility to the file system calls for the usage of the exclusively designed request control message. The blockchain users initiate the uploading process upon receiving the request control message from the comprehensive storage network. Similarly, the retrieval process also demands the usage of different variants of request control messages by the blockchain user.

This operation further facilitates accessing and storing the data on the user's device. This data could also be accessed (after authorization) by other legitimate users and administrators to obtain prior and new information. The proposed scheme constructs the entire network in a peer-to-peer method where the role of the secure storage service provider can be played by the user, who is further required to get registered to play the role of blockchain user. Upon accomplishment of this initial step, such a user can facilitate all the computing nodes and accessibility towards storage services. To eliminate the constraints associated with storage and channel capacity, the proposed system adopts using distributed cloud server as the primary point of storage instead of opting for multiple local storage units. Once the request control message is obtained for prompting towards file storage task, the proposed scheme applies asymmetric encryption for its data to be stored in the blockchain network. Therefore, adopting encryption over securing the primary data acts as a security shield against any attempt of a security breach.

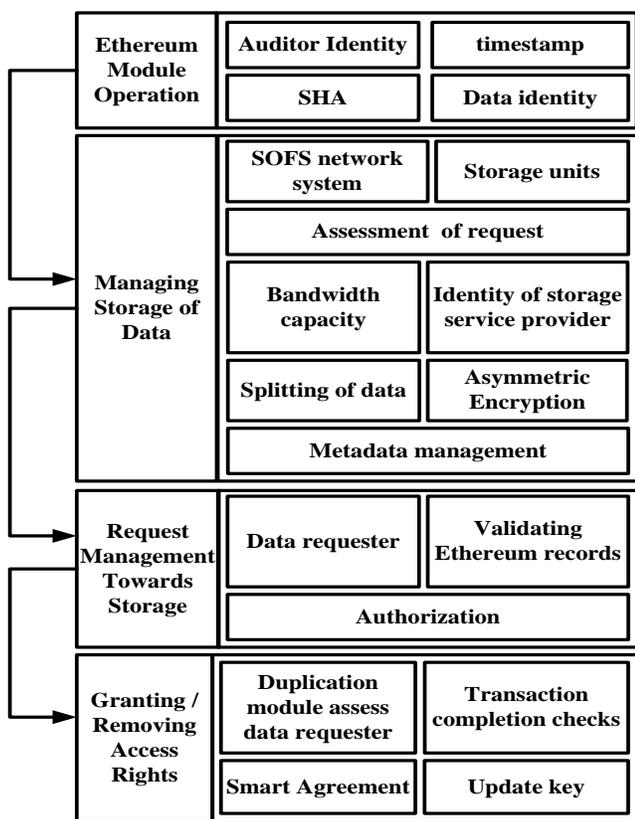


Fig. 1. Proposed architecture

Further the blockchain network is introduced with the recently encrypted data that also carry out sharing of information (related to secret key) in the blockchain network. This process also transforms private and public keys, which facilitates the blockchain user to gain potential control over the user's data either by revoking the request or permitting access. It should be noted that the proposed blockchain network performs interaction with the service provider of data using web 3.0. The usage of the Ethereum blockchain assists in managing multiple transactions records more effectively and securely. High-level planning of the functional framework

model is shown in Fig. 1 with the inclusion of the stakeholders in general, which can be easily customized to any functional domain along with the intrinsic technologies used for the security and privacy-preserving way data is stored and shared. According to Fig. 1, four essential blocks of operation are carried out towards achieving the target of privacy preservation, mainly emphasizing balancing the security demands and optimal computational efficiency. Each modular block carries a discrete set of interconnected operations where the core image file is subjected to security processing over the Ethereum distributed data system.

The complete operation stated in Fig. 1 is highly sequential from top to bottom. Initially, the Ethereum module is constructed with hash-based encryption followed by asymmetric encryption over the split data. A scheme also maintains a better indexing process using metadata management based on every request. The module also assists in performing validation of the Ethereum records followed by final management of removing or accessing rights permission. A unique and smart agreement is designed, followed by updating key. The core logic of the architecture mentioned above is to ensure faster and safer data repositioning over a cloud environment in a distributed manner. The consecutive section further illustrates all the blocks discussed earlier regarding algorithm design.

## V. ALGORITHM IMPLEMENTATION

From the discussion carried out in prior section, it is noted that proposed Ethereum-based ecosystem introduces a novel mechanism of image splitting which supports an extensive decentralization scheme followed by efficient request management for controlling image sharing. This section discusses the design of an algorithm implementation towards the methodology briefed in the prior section.

### A. Strategies for Algorithm Implementation

The algorithm's execution begins with a user who forwards its data to the interface, delivering the data to the request's handler. This request is sent to the Ethereum module, further interacting with the SOFS and duplication modules. In parallel with this process, the encrypted data from the storage is also forwarded to the Ethereum client. A decentralized datastore system using hashing is used to arrange this data. This hashed data is then sent to the transaction pool and authenticator module, further updated towards the Ethereum client. The data from the duplication module is now forwarded to the intelligent agreement management module controlled by the data owner. It is also delivered to the request handler for the next Ethereum client module. Updating operations of these transactions are simultaneously carried out in this process. Apart from this, the complete blocks of the data considered for the proposed implementation are the access scheme, duplication module, and pairing of keys. All these three blocks consist of essential information associated with the identity of the auditor for the Ethereum client, innovative agreement, the signature of the user, identity of the client, identity of data, value of hashed data, and time stamp. These are also the seven essential components of the Ethereum module. It should be noted that proposed scheme of privacy preservation adopted are hybridized form of both transactional and smart contract-based

privacy in Ethereum in order to keep a balance in storage and security at same time.

### B. Algorithm Design

The core algorithm of the proposed scheme consists of the following operations, i.e., Ethereum module operation, managing storage of data, request management towards storage, and granting/removing access rights. The core steps of the proposed algorithm are as follows:

Algorithm for Securing image using Ethereum
<b>Input:</b> $m$ (Components of Ethereum)
<b>Output:</b> $d_{val}$ (validated data)
<b>Start</b>
1. <b>For</b> $c=1:m$
2. $R_{req} \rightarrow E_c$
3. $E_c \rightarrow S_{net} \rightarrow$
4. $S_{net} \rightarrow val(avail(S_c))$
5. $S_{net} \rightarrow confirm(S_c)$
6. $data_1 \leftarrow f_1(Data)^n$
7. $data_2 \leftarrow f_2(data_1)$
8. $data_3 \rightarrow f_3(data_2)$
9.     obtain $d_{val}$
10. <b>End</b>
<b>End</b>

The discussion of prime operations toward each algorithmic step is as follows:

- *Ethereum Module Operation:* This is the initial module of implementation, which models the proposed Ethereum module focusing on privacy preservation towards image data. The proposed study considers the auditors' identity based on their computing devices. The study also uses a conventional Secured Hash Algorithm of a crucial size of 256 bits to hash the data for each image data chunk. A unique identifier is used for allocating the identity of stakeholders. There must be trust established between the SOFS network and the requestor node. The proposed system considers an innovative agreement, similar to smart contract, a well-structured program executed over the Ethereum module. Its primary responsibility is to validate all incoming data access requests for storage that are accessible to all the users of the Ethereum module. All the transactions of any form are carried out by the Ethereum module, which is highly distributed to offer better and faster access to files. The algorithm considers  $m$  number of components of the  $E_c$  Ethereum module (Line-1), where the data provider performs initiation of its reposition request  $R_{req}$  via its interface (Line-2).
- *Managing Storage of Data:* This module is responsible for managing all forms of incoming requests, encryption, and storage simultaneously. The Ethereum client  $E_c$  forwards its information to the  $S_{net}$  SOFS network (Line-3). It should be noted that all form of file request for storage is initially accepted by the storage units, which are regular users and part of the Ethereum client system interconnected via the cloud ecosystem.

The confirmation of the storage space  $S_c$  must be done by the Ethereum client (Line-5) by priorly validating them using the *val* method concerning their availability (Line-4). The  $S_c$  module carries out the request verification by constructing an array that retains all transactions of accepted files to be stored (Line-5). This operation significantly reduces time complexity from both storage and query processing. The  $S_c$  module carries out the identity of the storage unit offered by any cloud service provider concerning channel capacity and storage space availability with a duplication module.

- *Request Management Towards Storage:* This module carries out specific operation steps before managing all forms of requests. The complete input of an image data is split into the equivalent size of blocks of  $n$  number using an explicit function  $f_1(x)$  to generate split data  $data_1$  (Line-6). This operation offers a beneficial solution for dealing with more extensive or high-dimensional datasets. It is to be noted that the process carried out by SOFS is entirely decentralized, and hence there is no event of failures in storing split data over multiple storage units over the cloud. Further, an explicit function  $f_2(x)$  is used for performing the RSA, which is stored in the form of tree-based networks over the storage units (Line-7). This operation results in multiple encrypted data  $data_2$  (Line-7). The time complexity problem further reduces as the complete encryption is carried out in parallel to all the chunks of the data.
- Apart from this, storing the stakeholder's metadata on the  $E_c$  module's block is essential, although they are stored in the SOFS network system for their original data input. The proposed scheme also audited this data at a specific periodic interval, forwarding the signature to the following consecutive data blocks. The  $E_c$  will synchronize all the generated encrypted blocks of data. An interesting point from a security perspective is that a tiny amendment being carried out on any one block will change the whole setup of blocks of data due to any malicious activity. Hence, even if one block of data is stolen or compromised, it will be useless for an attacker. An additional layer of security is further implemented by storing the hashed value of these data to mitigate the problem of data leakage over cloud servers. The complete management of the data blocks is carried out so that the proposed algorithm evaluates the legitimacy of the data requester as a mandatory step. Suppose the requestor's identity is legitimate (from the metadata). In that case, it is added to the Ethereum distributed data system record, which is finally forwarded to the data storage unit. This operation potential assists in retaining maximum accountability of image data on SOFS network. Moreover, this accountability is carried out at period interval of time to keep the blocks updated thereby maintaining higher transparency. All the accessing of data can be carried out from this module. However, if the data requestor is found illegitimate (not a data owner), the proposed algorithm checks the access policy by updating the

records maintained by the Ethereum-distributed data system. In short, the Ethereum client maintains all the forms of the legitimate and illegitimate list of requestor nodes, resisting any attacker from accessing the file maliciously.

- *Granting / Removing Access Rights:* Using the verification process managed by the Ethereum client system, the valid data requestor can initiate the transactions, followed by regular updating towards the Ethereum client. However, for optimal security, the algorithm further steps towards updating its public or private key by generating a new version of it. The SOFS network system carries out this operation to control crucial sharing by granting or removing access rights. The beneficial aspect of this operation is that it can offer a higher degree of data privacy, eliminating the probability of key-based attacks. In this process, an explicit function  $f_3(x)$  is designed to manage the intelligent agreement system which is primarily response for handling data access request along with permission to be granted for access/deny (Line-8). For this purpose, the proposed scheme develops a duplication module for testifying its evidence based on four types of keys, i.e., keys to be used for preliminary instance, keys that are eliminated for the first time, and second and third time considering public and private keys. This operation generates  $data_3$ , which ultimately yields validated data  $d_{val}$  (Line-9). The proposed algorithm also maintains a record of all the access  $Rec$ . The process carried out by function  $f_3(x)$  can now be further extended: The algorithm assesses the event of successful completion of the accessing by the stakeholder, followed by eliminating the keys for all the identified instances. The intelligent agreement module reviews the complete record of access  $Rec$ , which finally generates a new record  $Rec$ . The algorithm enforces the stakeholder to request access if their old access record is not found in this  $Rec$ . The unique access is generated by yielding new key pairs, then applying asymmetric encryption to all the split data blocks and generating a key for encryption. The generation of the unique access is the contribution towards deploying a unique security measure to resist unauthorized data access. The storage system of the SOFS network receives this encoded file as well as the private key of the stakeholder. The file is further encrypted when it arrives within the SOFS network system while the user's private key is shared. All the keys are eliminated once the system records access completion, which makes the scheme high-level secured from intrusive activity on stored data.

## VI. RESULT ANALYSIS

The implementation of the proposed study is carried out on a conventional windows machine with 16 GB RAM and Core-i7 processor. An open-source server environment has been adopted for Java scripting the proposed algorithm. The proposed scheme also uses a high-level object-oriented language to deploy innovative agreements over an Ethereum environment. A standard benchmark test environment of

Mocha 6.2.0 is used, while Ganache is considered for the Ethereum platform. The complete assessment is carried out over a standard Kovan testnet. The standard nodes from Amazon Web Services, i.e., AWS nano, are considered for the storage nodes that perform their data transmission over 100mps with NVIDIA GEFORCE GTX as GPU. The performance parameters considered for assessing the proposed system with the existing system (conventional blockchain) are as follows:

- *Time for Data Reposition:* This is the time required to store the data in the distributed cloud servers after being subjected to block operations.
- *Time for Data Retrieval:* This is the time required for the stakeholder to access their stored data from distributed location to their local system.
- *Operational Cost:* This is the cumulative number of resources (in memory and bandwidth) used for overall repositioning and retrieval processing.
- *Delay:* This is latency associated with data transmission from one point to another. The study considers cumulative delay for both repositioning and retrieval.

The proposed system is evaluated based on 100 GB of sample data programmatically generated traffic in IoT. Assessed over 1200 simulation rounds, the sample data is allocated and increased arbitrarily to map with a practical world environment.

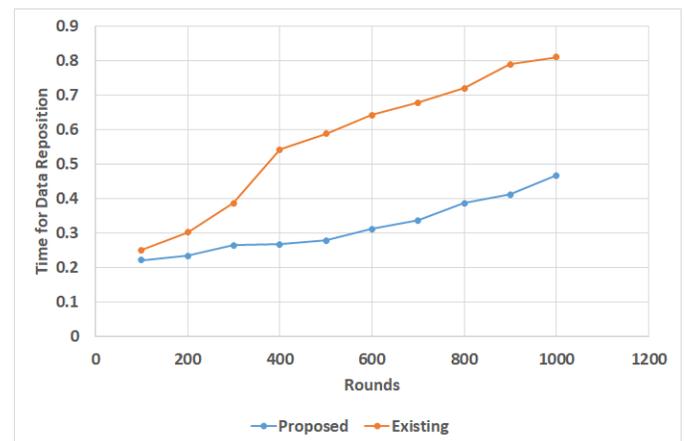


Fig. 2. Analysis of time for data reposition

Fig. 2 highlights the time required to store the data, stating the proposed system offers very little time consumption compared to conventional blockchain technology. The prime reason behind this is the higher dependency on adopting consensus-based mechanisms by traditional blockchain technology, which also induces scalability issues. On the other hand, the proposed scheme doesn't have any such dependencies that result in faster processing. Apart from this, conventional blockchain technology is inherently characterized by a slower process if the size of the network increases. A closer look into the proposed scheme shows that the tree-based structuring of hashed data makes the flow of the encoded blocks of spitted data much faster than the existing scheme.

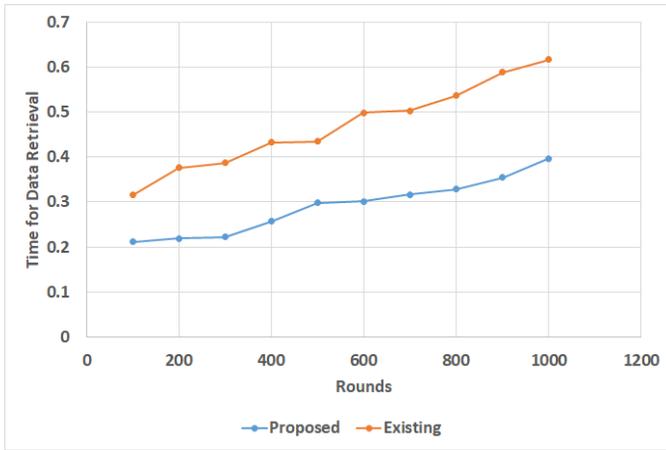


Fig. 3. Analysis of time for data retrieval

Fig. 3 highlights that the proposed system offers much-reduced consumption of time for retrieval of data in comparison to the conventional blockchain. A similar reason for time for repositioning can be stated as its cause. Apart from this, the process of metadata management by the SOFS network system makes the faster process of requestor legitimacy. Although it depends upon the synchronization time of updating, it still offers speedier query processing. Conventional blockchain has a higher dependency on node operation. In contrast, the proposed scheme outsources this dependency towards the Ethereum distributed database system using the SOFS network system, making the retrieval system relatively faster with a speedy auditing process for the information request.

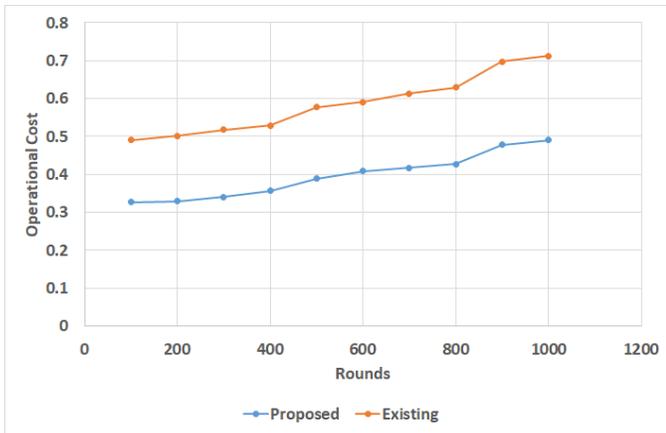


Fig. 4. Analysis of operational cost

Fig. 4 showcases that the proposed system offers reduced operational costs compared to conventional blockchain technology. Operational cost is one of the prime performance parameters to ascertain the applicability of the blockchain process toward data security in a realistic environment. The prime rationale behind this outcome is that traditional blockchain enforces the miners to solve problems with updates of new transactions by the ledger, which consistently increases resource dependencies. On the contrary, the proposed scheme controls its resource inclusion as a preemptive method of computing the entire path of data forwarding to the distributed

storage unit, considering all the constraints from a security perspective. This offers a reduced consumption of resources leading to reduced operational costs.

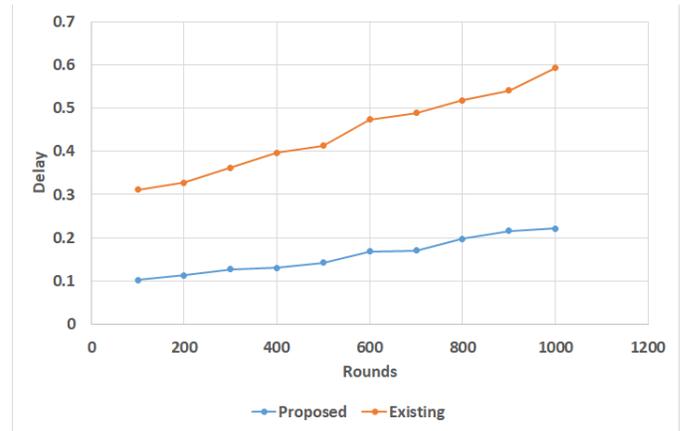


Fig. 5. Analysis of delay

Fig. 5 showcases that the proposed system offers a significantly reduced delay trend compared to conventional blockchain. It is to be noted that traditional blockchain forces the individual to have their key to make it completely decentralized. This process offers a sophisticated key management scheme and increases the information stored and retrieved. This eventually leads to a potential lag in time. However, the proposed system provides innovative agreement management, supporting faster auditing tasks over split encrypted data. Hence, irrespective of the usage of the RSA scheme, there are no complications towards key management both during storing and retrieval. The proposed system offers reduced delay trends to support various online applications over distributed cloud applications and services. Hence, the proposed method provides evidence to show its complete control over reducing time complexity.

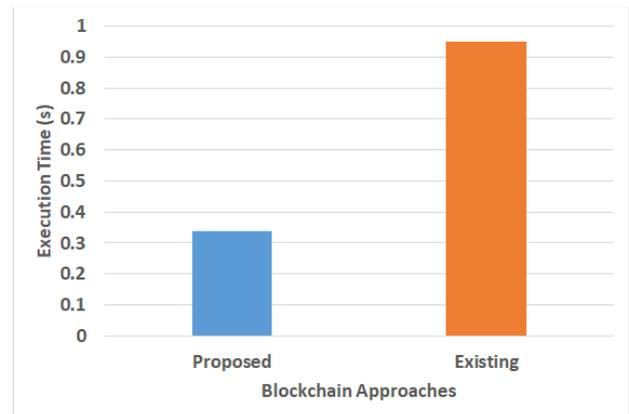


Fig. 6. Analysis of execution time

From the point of benchmarking, the proposed evaluation process considers assessing execution time. It should be noted that an indirect evaluation of memory is analyzed in the operational cost metric exhibited in Fig. 4. The outcome exhibited in Fig. 6 showcases that the proposed blockchain approach offers approximately 61% faster execution time compared to the existing blockchain scheme. A similar reason

stated for preliminary analysis is also to be attributed for this outcome. Hence, it is found that the proposed system can perform cost-effective computational operations in IoT.

## VII. DISCUSSION

The graphical outcome highlighted in the prior section shows that the proposed blockchain-based operation offers better performance than conventional blockchain architecture in various respects. One of the prime reasons for the betterment in performance is its non-dependency from any entities associated with a trusted third party as witnessed in existing system [11]. This process is enabled using the proposed smart agreement system incorporated in the system design connected to processing the data over the network. This fact also contributes towards a lesser operational cost score for the proposed system, as noted in Fig. 4. The outcome will also contribute towards an effective large-scale data-sharing process among multiple enterprises. Apart from this, it can be noted that the proposed scheme offers reduced delay (Fig. 5) and time of retrieval (Fig. 3); owing to faster computational speed, the proposed scheme significantly controls the scalability problem in conventional blockchain design. This process is implemented by using the sharding mechanism.

From the perspective of the privacy preservation viewpoint, there is a potential difference between conventional and proposed Ethereum design methodology. The conventional Ethereum experience low processing speed whereas processing speed is much lower for proposed scheme. Apart from this, proposed scheme maintains a novel blockchain network topology with higher control of data and transaction where the compliance is monitored using smart agreement with updates. This offers higher privacy conservation in contrast to existing Ethereum.

Unlike the existing mechanism of secure data sharing [7][8][12][13][16][25], as illustrated in Section II, the proposed system enables the direct accessibility of the data to its owner. The data owner can carry out the associated process of its access policy and its need for amendment or customization. Hence, complete control of data ownership is retained in the proposed scheme in adherence with the smart agreement of laws towards privacy protection. From the perspective of legal authority of General Data Protection Regulation (GDPR), it is known that blockchain system is always considered to possess a data controller (where at least one legal person resides) in order to ensure the correct implementation of law of data protection in EU. Such forms of data controllers are mandatory require to adhere to the protocols of GDPR. Hence, the proposed scheme of Ethereum, in spite of its decentralized scheme, always offers its architecture to be controlled by multiple authorities in order to ensure data privacy as per the algorithm implementation.

## VIII. CONCLUSION

The proposed concept presents a discussion about a novel computational framework that harnesses the potential of the Ethereum distributed database system for facilitating a secure validation of the data or any participating nodes towards storing and retrieving the data from distributed cloud servers. The novelty of the proposed scheme is as follows:

1. The model presents a duplication module for evidence to offer an assurance of the legitimacy of each transaction being carried out by the Ethereum-based distributed database system
2. The proposed scheme introduces a novel image data-sharing process and associated access policy management using a unique intelligent agreement system.
3. The complete assessment is carried out over Amazon Web Service nano nodes for standardizing its outcome using the SOFS network system and Ethereum client
4. The proposed scheme can control all the problems that impede decentralization using distributed storage units organized by the SOFS network system. At the same time, the complexity associated with data security and scalability is addressed using the presented splitting of data followed by encryption on every split data.
5. The proposed system offers approximately i) 42.5% of reduction in time for data repositioning, ii) 41.1% of reduction in time for data retrieval, iii) 34.8% of reduction in operational cost, and iv) 73.9% of reduction in delay trend in comparison to conventional blockchain method.

The possible shortcoming of the paper is that it is further required to be evaluated on real ground and find a similar consistency in its outcome. More test environments are further required to ensure this. From the above outcomes assessed on a standard benchmarked testbed, it can be stated that the proposed scheme offers better control over time complexity and high-end data privacy. Future work will further extend the present model toward more optimization-based processing. For this purpose, various bio-inspired approaches will be investigated to improve the blockchain operation further. The major emphasis will also be given to the impact of the massive peak and concurrent bottleneck conditions on the performance of blockchain operations. Further in order to offer higher privacy preservation over proposed scheme, the future work direction will be to extract the stochastic trends of dynamic attacker considering the network attributes to develop a novel attack map. This distributed attack map can be used for sandboxing any form of illegitimate or suspicious data request to further confirm the legitimacy of the request. Improving upon encryption protocol over such distributed attack map is anticipated to offer higher degree of privacy preservation.

## REFERENCES

- [1] H. Li, J. Liu, and X. Zhou, "Intelligent Map Reader: A Framework for Topographic Map Understanding With Deep Learning and Gazetteer," in *IEEE Access*, vol. 6, pp. 25363-25376, 2018, doi: 10.1109/ACCESS.2018.2823501
- [2] Y. Zheng et al., "Histopathological Whole Slide Image Analysis Using Context-Based CBIR," in *IEEE Transactions on Medical Imaging*, vol. 37, no. 7, pp. 1641-1652, July 2018, doi: 10.1109/TMI.2018.2796130
- [3] I. Z. Hong, D. Ming, K. Zhou, Y. Guo, and T. Lu, "Road Extraction From a High Spatial Resolution Remote Sensing Image Based on Richer Convolutional Features," in *IEEE Access*, vol. 6, pp. 46988-47000, 2018, doi: 10.1109/ACCESS.2018.2867210.
- [4] B. Jiang, M. Seif, R. Tandon, and M. Li, "Context-Aware Local Information Privacy," in *IEEE Transactions on Information Forensics*

- and Security, vol. 16, pp. 3694-3708, 2021, doi: 10.1109/TIFS.2021.3087350.
- [5] J. H. Abawajy, M. I. H. Ninggal, and T. Herawan, "Privacy Preserving Social Network Data Publication," in IEEE Communications Surveys & Tutorials, vol. 18, no. 3, pp. 1974-1997, third quarter 2016, doi: 10.1109/COMST.2016.2533668.
- [6] A. K. Sandhu, "Big data with cloud computing: Discussions and challenges," in Big Data Mining and Analytics, vol. 5, no. 1, pp. 32-40, March 2022.
- [7] G. Akkuzu, B. Aziz and M. Adda, "Towards Consensus-Based Group Decision Making for Co-Owned Data Sharing in Online Social Networks," in IEEE Access, vol. 8, pp. 91311-91325, 2020, doi: 10.1109/ACCESS.2020.2994408.
- [8] L. Dong et al., "A Hierarchical Distributed Processing Framework for Big Image Data," in IEEE Transactions on Big Data, vol. 2, no. 4, pp. 297-309, 1 December 2016, doi: 10.1109/TBDATA.2016.2613992.
- [9] R. G. Sonkamble, S. P. Phansalkar, V. M. Potdar, and A. M. Bongale, "Survey of Interoperability in Electronic Health Records Management and Proposed Blockchain Based Framework: MyBlockEHR," in IEEE Access, vol. 9, pp. 158367-158401, 2021, doi: 10.1109/ACCESS.2021.3129284.
- [10] S. Bhatt, T. K. Pham, M. Gupta, J. Benson, J. Park, and R. Sandhu, "Attribute-Based Access Control for AWS Internet of Things and Secure Industries of the Future," in IEEE Access, vol. 9, pp. 107200-107223, 2021, doi: 10.1109/ACCESS.2021.3101218.
- [11] W. Fang, N. Cui, W. Chen, W. Zhang, and Y. Chen, "A Trust-Based Security System for Data Collection in Smart City," in IEEE Transactions on Industrial Informatics, vol. 17, no. 6, pp. 4131-4140, June 2021, doi: 10.1109/TII.2020.3006137.
- [12] H. R. Hasan et al., "A Blockchain-Based Approach for the Creation of Digital Twins," in IEEE Access, vol. 8, pp. 34113-34126, 2020, doi: 10.1109/ACCESS.2020.2974810.
- [13] M. Zichichi, S. Ferretti and G. D'Angelo, "A Framework Based on Distributed Ledger Technologies for Data Management and Services in Intelligent Transportation Systems," in IEEE Access, vol. 8, pp. 100384-100402, 2020, doi: 10.1109/ACCESS.2020.2998012.
- [14] S. Jiang and J. Wu, "A Blockchain-Powered Data Market for Multi-User Cooperative Search," in IEEE Transactions on Network and Service Management, vol. 19, no. 1, pp. 203-215, March 2022. doi: 10.1109/TNSM.2021.3125604
- [15] M. Debe, K. Salah, M. H. Ur Rehman and D. Svetinovic, "Monetization of Services Provided by Public Fog Nodes Using Blockchain and Smart Contracts," in IEEE Access, vol. 8, pp. 20118-20128, 2020. doi: 10.1109/ACCESS.2020.2968573
- [16] H. R. Hasan, K. Salah, R. Jayaraman, I. Yaqoob, M. Omar and S. Ellahham, "Blockchain-Enabled Telehealth Services Using Smart - Contracts," in IEEE Access, vol. 9, pp. 151944-151959, 2021. doi: 10.1109/ACCESS.2021.3126025
- [17] B. Podgorelec, M. Heričko and M. Turkanović, "State Channel as a Service Based on a Distributed and Decentralized Web," in IEEE Access, vol. 8, pp. 64678-64691, 2020. doi: 10.1109/ACCESS.2020.2984378
- [18] [18] M. M. Madine et al., "Fully Decentralized Multi-Party Consent Management for Secure Sharing of Patient Health Records," in IEEE Access, vol. 8, pp. 225777-225791, 2020. doi: 10.1109/ACCESS.2020.3045048
- [19] E. M. Abou-Nassar, A. M. Ilyasu, P. M. El-Kafrawy, O. -Y. Song, A. K. Bashir, and A. A. A. El-Latif, "DITrust Chain: Towards Blockchain-Based Trust Models for Sustainable Healthcare IoT Systems," in IEEE Access, vol. 8, pp. 111223-111238, 2020. doi: 10.1109/ACCESS.2020.2999468
- [20] C. Lin, D. He, X. Huang, X. Xie, and K. -K. R. Choo, "PPChain: A Privacy-Preserving Permissioned Blockchain Architecture for Cryptocurrency and Other Regulated Applications," in IEEE Systems Journal, vol. 15, no. 3, pp. 4367-4378, Sept. 2021. doi: 10.1109/JSYST.2020.3019923
- [21] A. Kumar, R. Krishnamurthi, A. Nayyar, K. Sharma, V. Grover, and E. Hossain, "A Novel Smart Healthcare Design, Simulation, and Implementation Using Healthcare 4.0 Processes," in IEEE Access, vol. 8, pp. 118433-118471, 2020. doi: 10.1109/ACCESS.2020.3004790
- [22] Z. Ullah, B. Raza, H. Shah, S. Khan, and A. Waheed, "Towards Blockchain-Based Secure Storage and Trusted Data Sharing Scheme for IoT Environment," in IEEE Access, vol. 10, pp. 36978-36994, 2022. doi: 10.1109/ACCESS.2022.3164081
- [23] X. Yan, X. Yuan, Q. Ye, and Y. Tang, "Blockchain-Based Searchable Encryption Scheme With Fair Payment," in IEEE Access, vol. 8, pp. 109687-109706, 2020. doi: 10.1109/ACCESS.2020.3002264
- [24] W. Xiong and L. Xiong, "Smart Contract Based Data Trading Mode Using Blockchain and Machine Learning," in IEEE Access, vol. 7, pp. 102331-102344, 2019. doi: 10.1109/ACCESS.2019.2928325
- [25] M. Debe, K. Salah, M. H. U. Rehman and D. Svetinovic, "IoT Public Fog Nodes Reputation System: A Decentralized Solution Using Ethereum Blockchain," in IEEE Access, vol. 7, pp. 178082-178093, 2019. doi: 10.1109/ACCESS.2019.2958355
- [26] H. R. Hasan, K. Salah, I. Yaqoob, R. Jayaraman, S. Pesic and M. Omar, "Trustworthy IoT Data Streaming Using Blockchain and IPFS," in IEEE Access, vol. 10, pp. 17707-17721, 2022. doi: 10.1109/ACCESS.2022.3149312
- [27] H. Yang, J. Yuan, H. Yao, Q. Yao, A. Yu, and J. Zhang, "Blockchain-Based Hierarchical Trust Networking for JointCloud," in IEEE Internet of Things Journal, vol. 7, no. 3, pp. 1667-1677, March 2020. doi: 10.1109/JIOT.2019.2961187
- [28] T. Chen et al., "GasChecker: Scalable Analysis for Discovering Gas-Inefficient Smart Contracts," in IEEE Transactions on Emerging Topics in Computing, vol. 9, no. 3, pp. 1433-1448, 1 July-Sept. 2021. doi: 10.1109/TETC.2020.2979019
- [29] Saini, Q. Zhu, N. Singh, Y. Xiang, L. Gao and Y. Zhang, "A Smart-Contract-Based Access Control Framework for Cloud Smart Healthcare System," in IEEE Internet of Things Journal, vol. 8, no. 7, pp. 5914-5925, 1 April 1, 2021. doi: 10.1109/JIOT.2020.3032997
- [30] M. Debe, K. Salah, M. H. U. Rehman and D. Svetinovic, "Blockchain-Based Decentralized Reverse Bidding in Fog Computing," in IEEE Access, vol. 8, pp. 81686-81697, 2020. doi: 10.1109/ACCESS.2020.2991261
- [31] B. Kaynak, S. Kaynak and Ö. Uygun, "Cloud Manufacturing Architecture Based on Public Blockchain Technology," in IEEE Access, vol. 8, pp. 2163-2177, 2020. doi: 10.1109/ACCESS.2019.2962232
- [32] Y. Zhang, C. Xu, J. Ni, H. Li, and X. S. Shen, "Blockchain-Assisted Public-Key Encryption with Keyword Search Against Keyword Guessing Attacks for Cloud Storage," in IEEE Transactions on Cloud Computing, vol. 9, no. 4, pp. 1335-1348, 1 Oct.-Dec. 2021. doi: 10.1109/TCC.2019.2923222
- [33] R. Akkoui, X. Hei and W. Cheng, "EdgeMediChain: A Hybrid Edge Blockchain-Based Framework for Health Data Exchange," in IEEE Access, vol. 8, pp. 113467-113486, 2020. doi: 10.1109/ACCESS.2020.3003575
- [34] S. Wang, R. Pei and Y. Zhang, "EIDM: A Ethereum-Based Cloud User Identity Management Protocol," in IEEE Access, vol. 7, pp. 115281-115291, 2019. doi: 10.1109/ACCESS.2019.2933989
- [35] O. Alkadi, N. Moustafa, B. Turnbull, and K. -K. R. Choo, "A Deep Blockchain Framework-Enabled Collaborative Intrusion Detection for Protecting IoT and Cloud Networks," in IEEE Internet of Things Journal, vol. 8, no. 12, pp. 9463-9472, 15 June 15, 2021. doi: 10.1109/JIOT.2020.2996590
- [36] S. Wang, X. Wang and Y. Zhang, "A Secure Cloud Storage Framework With Access Control Based on Blockchain," in IEEE Access, vol. 7, pp. 112713-112725, 2019. doi: 10.1109/ACCESS.2019.2929205
- [37] Y. Zhang, R. H. Deng, X. Liu, and D. Zheng, "Outsourcing Service Fair Payment Based on Blockchain and Its Applications in Cloud Computing," in IEEE Transactions on Services Computing, vol. 14, no. 4, pp. 1152-1166, 1 July-Aug. 2021. doi: 10.1109/TSC.2018.2864191
- [38] Y. Yang, H. Lin, X. Liu, W. Guo, X. Zheng, and Z. Liu, "Blockchain-Based Verifiable Multi-Keyword Ranked Search on Encrypted Cloud With Fair Payment," in IEEE Access, vol. 7, pp. 140818-140832, 2019. doi: 10.1109/ACCESS.2019.2943356
- [39] D. C. Nguyen, P. N. Pathirana, M. Ding and A. Seneviratne, "Blockchain for Secure EHRs Sharing of Mobile Cloud Based E-Health

- Systems," in *IEEE Access*, vol. 7, pp. 66792-66806, 2019. doi: 10.1109/ACCESS.2019.2917555
- [40] Y. Wang, A. Zhang, P. Zhang and H. Wang, "Cloud-Assisted EHR Sharing With Security and Privacy Preservation via Consortium Blockchain," in *IEEE Access*, vol. 7, pp. 136704-136719, 2019. doi: 10.1109/ACCESS.2019.2943153
- [41] X. Zhang, J. Zhao, C. Xu, H. Li, H. Wang, and Y. Zhang, "CIPPPA: Conditional Identity Privacy-Preserving Public Auditing for Cloud-Based WBANs Against Malicious Auditors," in *IEEE Transactions on Cloud Computing*, vol. 9, no. 4, pp. 1362-1375, 1 Oct.-Dec. 2021. doi: 10.1109/TCC.2019.2927219
- [42] R. Awadallah, A. Samsudin, J. S. Teh, and M. Almazrooie, "An Integrated Architecture for Maintaining Security in Cloud Computing Based on Blockchain," in *IEEE Access*, vol. 9, pp. 69513-69526, 2021. doi: 10.1109/ACCESS.2021.3077123.

# Music Note Feature Recognition Method based on Hilbert Space Method Fused with Partial Differential Equations

Liqin Liu

Music School, Hubei Engineering University, Hubei, 432000, China

**Abstract**—Hilbert space method is an old mathematical theoretical model developed based on linear algebra and has a high theoretical value and practical application. The basic idea of the Hilbert space method is to use the existence of some stable relationship between variables and to use the dynamic dependence between variables to construct the solution of differential equations, thus transforming mathematical problems into algebraic problems. This paper firstly studies the denoising model in the process of music note feature recognition based on partial differential equations, then analyzes the denoising method based on partial differential equations and gives an algorithm for fused music note feature recognition in Hilbert space; secondly, this paper studies the commonly used music note feature recognition methods, including linear predictive cepstral coefficients, Mel frequency cepstral coefficients, wavelet transform-based feature extraction methods and Hilbert space-based feature extraction methods. Their corresponding feature extraction processes are given.

**Keywords**—Partial differential equation; Hilbert space method; musical note feature recognition method; cepstral coefficients; empirical modal

## I. INTRODUCTION

With the continuous progress of science and technology, the development of high and new technologies such as computer technology, information technology, and microelectronics, and the continuous improvement of computer computing power, people are paying more and more attention to how to find the objective function in complex problems that can be solved by lower-order variables [1]. The traditional method can no longer meet the higher precision requirements in researching mathematical problems, and it is increasingly challenging to meet practical application requirements. Traditional methods are generally studied by analyzing variables and constructing a step response function [2]. Usually, each parameter in the differential equation needs to be derived, and the corresponding equation is obtained by using the variation relationship of partial derivatives of each variable at different orders. The Hilbert space is composed of many independent variables, and these independent variables have their corresponding equation expressions at different orders [3].

The traditional denoising method of music noise is a point-by-point iterative method with good timeliness. However, many iterative calculations and numerical analyses are often required to establish partial differential equations, which are computationally intensive, time-consuming, and challenging to

obtain accurate results [4]. This paper establishes a numerical music noise model based on partial differential equations, and the initial value problem of partial differential equations is transformed into numerical and iterative noise. The model is simple, convenient, easy to implement, and accurate. In this method, the continuous function in the time domain is first transformed into a higher-order system state equation, then converted into a discrete form through a series of inverse transformations, and then processed with partial derivatives, thereby realizing the numerical analysis of partial differential equations [5]. This method can effectively solve many nonlinear problems, and the calculation accuracy is improved to a certain extent compared with other algorithms.

The traditional signal analysis techniques commonly used at present are based on linear predictive analysis or Fourier analysis, or Wavelet analysis techniques. The principle of these techniques is based on processing raw data, decomposing it into a linear model, and then describing it with Fourier transform [6]. Therefore, these techniques are mainly aimed at analyzing linear and stationary signals, but for nonlinear signals, the linear model is often difficult to describe effectively, and the Hilbert space method can make good use of these theories [7]. The Hilbert space method is a new nonlinear signal feature recognition method proposed by Academician Huang E et al. The principle of this method is based on the decomposition of the Hilbert space method. The SCV or FFTM technique synthesizes a linear stationary non-second-order derivative sequence signal to represent its characteristics. The Hilbert space method can be used to identify nonlinear signals and effectively applied to identify linear signals, which has high practical application value.

## II. RELATED WORK

In this case, we use the partial differential equation learning model to solve the challenging facial recognition problem. There is a proposal for a unique feature selection method that uses a learning model based on partial differential equations. Because of this, the extracted features are more resistant to shifts in lighting conditions and can be rotated and translated without losing their integrity. This article by Xia Miao (2021) employs the face detection algorithm in face recognition technology to first detect the face and intercept the expression data, then calculates the increase rate, all in gauging students' focus in class. The expression is then scored based on the revised model of concentration analysis and evaluation of a college Chinese class, which is utilized to identify the

expression. In the end, the concentration score is the expression score multiplied by the head-up rate. Experiments are conducted in real classrooms, and findings are analyzed to draw appropriate conclusions and instructional recommendations [8]. The sparse representation of sample points in the neighborhood is created after the k-nearest neighbor approach selects a large neighborhood set for each face, thus combining the locality of the k-nearest neighbor with the robustness of sparse representation. Utilizing sparse reconstruction coefficients to characterize neighborhood geometry and weighted distance to characterize class dissimilarity, the sparse preserving nonnegative block alignment approach builds a discriminant partial optimization model. The two algorithms are successful and robust, as evidenced by their ability to produce accurate clustering and recognition results across a wide range of conditions, including both real and simulated occlusion. This study validated the model through in-class practice assessments, teachers' inquiries, and interviews with students and teachers to ensure its accuracy. The outcomes demonstrate the validity and trustworthiness of the proposed combined evaluation approach based on expression and head-up rate.

As a preprocessing step in a wide variety of applications, such as sound separation and musical note transcription, musical pitch estimation is used to identify the pitch of a musical note or the fundamental frequency (F0) of an audio stream. Based on the categorization framework, Tamboli (2019) creates a neural network optimized for musical note recognition (OBNN). The strategies for identifying musical notes were identified after reviewing a variety of surveys and studies. Here, an OBNN is utilized to identify musical pitches. Similarly, by utilizing various approaches, we can improve the efficiency of musical note recognition [9]. The most recent studies on musical note identification are successfully summarized here, along with the characteristics and categorizations gleaned from those studies.

Automatic speech/music classification employs various signal processing methods to sort audio/visual files into predefined categories. In order to categorize incoming audio signals into speech/music signals, Arvind Kumar's (2022) suggested work investigates Hilbert Spectrum (HS) obtained from various AM-FM components of an audio signal, also known as Intrinsic Mode Functions (IMFs). Hilbert Transform of the IMFs yields a two-dimensional representation of the HS, a map of instantaneous energy (IE) and frequencies (IF). Via creating unique IF and Instantaneous Amplitude (IA) based cepstral features, we subject this HS to a Mel-filter bank and Discrete Cosine Transform (DCT). Three datasets (Slaney Database, GTZAN Database, and MUSAN Database) were used to validate the results. Extensive experiments were undertaken on various combinations of audio files from the S&S, GTZAN, and MUSAN databases to evaluate the broad applicability of the proposed characteristics, and positive results were attained. Finally, the system's performance is compared to previously implemented cepstral features and other related efforts [10].

Integrating AI with deep music for recommendations has been a growing area of study in recent years. Deep learning is a complicated machine learning technique that may infer value

laws from features observed in training samples. The proliferation of deep learning networks is key to the future of AI and offers a fresh perspective on music score identification. Qin Lin's (2022) paper utilizes the enhanced deep learning algorithm to study music score recognition. To achieve feature extraction and intelligent recognition of music scores, we build on the foundation of the classic neural network by introducing the attention weight value improved convolutional neural network (CNN) and the high execution efficiency deep belief network (DBN). A CNN&DBN-based feature learning method was developed for music score extraction using the feature vector set extracted by CNN&DBN as input [11]. Experimental results show that the proposed model effectively recognizes a wide range of polyphony music types, with improved recognition and performance; the improved algorithm applied to soundtrack identification achieves a recognition rate of up to 98.4%, which is significantly higher than those of other classic algorithms. It demonstrates the massive potential for study in the field of music retrieval using deep learning and provides data support for building a knowledge graph in the music field.

It is a common goal in engineering and computer science to give machines sensing abilities on par with those of humans. Much work has been done to give computers the ability to collect, process, evaluate, and understand their environment in the same ways humans do. Explicitly referring to the auditory system, machine hearing is the capacity of computers to perceive their acoustic surroundings in the same way humans do. A proper audio signal representation is crucial to accomplishing this lofty goal. This study by Alas F et al. (2016) provides a comprehensive overview of the most recent advances in audio feature extraction methods for analyzing standard audio signals like voice, music, and environmental noise [12]. For the sake of thoroughness, the writers revisited old methods and included the most recent developments based on new fields of research and unique bio-inspired recommendations. These methods are classified in a taxonomy that groups them by their physical or perceptual underpinnings and then further subdivided by the type of computing they do (time, frequency, wavelet, image-based, cepstral, or other domains). The methods are described, and recent applications to issues with machine hearing are provided as illustrative instances.

Since its introduction, the Hilbert-Huang transform method has been widely used thanks to its superiority in several different contexts. The Hilbert spectrum accurately reflects the signal energy's dispersion over multiple scales. Using the Hilbert energy spectrum, which characterizes the distribution of instantaneous energy, Li X (2011) proposes a novel characteristic dubbed ECC. Compared to the conventional short-term average energy, the experimental findings clearly show that ECC performs better [13]. Combining the ECC and mel frequency cepstral coefficients (MFCC) provides a more detailed picture of the energy distribution over both the temporal and frequency domains, and the features of this set outperform those of the short-term average energy, pitch, and MFCC in terms of recognition accuracy. Then, after that, new and improved ECC variants are created. Combine ECC with the teager energy operator to get TECC, and add the

instantaneous frequency to the energy to get EFCC. Seven different emotional states are tested, with boredom having the highest detection rate (83.57%) and the highest categorization accuracy (100%). The proposed characteristics ECC, TECC, and EFCC were shown to enhance speech emotion recognition performance in numerical tests significantly.

Film video noise is commonly understood as digital signal system errors manifested as artifacts in the video image. Videos captured with different cameras will always have some degree of this distortion. The primary purpose of noise reduction is to lessen the amount of distracting background noise in a video while allowing the image's edges and textures to come through clearly. Pingli Sun et al. (2021) provide a comprehensive explanation of the space-time noise reduction filter's workings, along with the development of a 3D-filter algorithm for Gaussian noise, an enhanced 3D-filter algorithm for mixed noise based on the 3D-BDP (bloom-deep-split) filter, and a filter algorithm for luminance and color noise in dimly lit scenes [14]. They build a novel iterative denoising algorithm by deconstructing the PDE denoising process. Partial differential equations can be thought of as an iterative denoising of the filter. The new algorithm's initial stage employs a wavelet-domain adaptive Wiener filter as its filtering foundation, with promising results achieved by careful tuning of the filter's parameters. Analysis findings demonstrate that the model proposed in this section can efficiently eliminate multiplicative noise compared to the existing denoising model. The experimental report demonstrates that, compared to the partial differential equation method for denoising, the algorithm's parameters have some stability and can obtain satisfactory processing outcomes for many images. Using the proper partial differential equation approach, the pseudo-Gibbs are eliminated in the second step of the algorithm, greatly enhancing its performance. After applying the new algorithm to a Gaussian-noise-filled image, the pseudo-Gibbs effect, which frequently occurs in wavelet denoising, and the step effect, which occurs in partial differential equation denoising, are both eliminated, details are better preserved, the peak signal-to-noise ratio is improved, and numerous experiments demonstrate the algorithm's efficacy as a denoising method.

### III. DENOISING MODEL OF MUSICAL NOTE FEATURES BASED ON PARTIAL DIFFERENTIAL EQUATIONS

The music note feature denoising method based on a partial differential equation mainly uses the variational denoising method (TV) to identify the noise feature. The principle of this method is to transform the denoising problem into an extreme problem of finding the energy functional expression established under two constraints. The basic idea of this method is to transform an extreme value problem with a time window constraint function into an unconstrained next-dimensional signal and solve it in the time domain, and then obtain the actual non-stationary random process through iterative transformation, and finally realize the estimation of the noise spectrum [15].

Before studying the de-algorithm, the noise reduction model needs to be studied first. The noise reduction model is expressed as formula (1):

$$u_o(x, y) = u(x, y) + n(x, y) \quad (1)$$

Based on the noise reduction model, related personnel proposed a global variational model based on a bounded variation space, and its definition is shown in formula (2):

$$TV(u) = \int_{\Omega} |\nabla u(x, y)| d\Omega \quad (2)$$

Formula (2) satisfies constraints (3), (4):

$$\int_{\Omega} u(x, y) dx dy = \int_{\Omega} u_o(x, y) dx dy \quad (3)$$

$$\int_{\Omega} \frac{1}{2} |u(x, y) - u_o(x, y)|^2 dx dy = \sigma^2 \quad (4)$$

Usually, the noise will make the overall variational energy of the acquired musical note features large, and it is difficult to identify accurately. The noise reduction model can optimize the overall variational model of musical note features, that is, to minimize the overall variational energy, and it can accurately identify the temporal and spatial dimensions of music. Therefore, the denoising of musical note features is mainly to minimize the energy functional, which can be expressed as formula (5) by using the Lagrange multiplier method:

$$\hat{u} = \arg \min_u \{ E(u) \} = \frac{\lambda}{2} \int_{\Omega} (u - u_o)^2 dx dy + \int_{\Omega} \sqrt{u_x^2 + u_y^2} dx dy \quad (5)$$

The Euler-Lagrange equation corresponding to formula (5) can be deduced by the variational method, as shown in formula (6):

$$\lambda(u - u_o) - \frac{\partial}{\partial x} \left( \frac{u_x}{\sqrt{u_x^2 + u_y^2}} \right) - \frac{\partial}{\partial y} \left( \frac{u_y}{\sqrt{u_x^2 + u_y^2}} \right) = 0 \quad (6)$$

The gradient descent flow equation corresponding to the variational problem Equation (6) is shown in Equation (7):

$$\begin{aligned} \frac{\partial u}{\partial t} &= \frac{\partial}{\partial x} \left( \frac{u_x}{\sqrt{u_x^2 + u_y^2}} \right) + \frac{\partial}{\partial y} \left( \frac{u_y}{\sqrt{u_x^2 + u_y^2}} \right) - \lambda(u - u_o) \\ &= \text{div} \left[ \frac{\nabla u}{|\nabla u|} \right] - \lambda(u - u_o) \end{aligned} \quad (7)$$

TV flow is stable and has a globally optimal solution. It is a diffusion function between forward diffusion and progressive diffusion. This function can obtain a better diffusion process and feature information effect of musical note features [16]. The numerical implementation of the TV model is as follows:

The numerical realization of the TV model usually has three differential schemes: explicit, implicit, and semi-implicit. The implicit and semi-implicit schemes have higher secrecy than the explicit scheme. However, for the semi-implicit and implicit numerical implementation methods, the secrecy makes the whole calculation process more complicated in the iterative process of removing noise. Complex and the convergence rate is relatively slow. Compared with the other two numerical iterative implementations, the numerical implementation using

clear difference has a much faster convergence rate [17]. Therefore, the display scheme is commonly used by people, and the implementation of the display scheme is as follows:

The first display scheme is available as shown in formula (8):

$$u^{n+1} = u^n + \frac{\nabla t}{2\lambda} \operatorname{div} \left( \frac{\nabla u^n}{|\nabla u^n|} \right) - \nabla t \frac{(u_0 - u^n)(\mu u_0 - \mu + 1)}{(\mu u^n - \mu + 1)^3} \quad (8)$$

Substituting the difference quotient for the partial derivative, the formula (9) can be obtained:

$$\begin{aligned} \nabla_{\pm}^x u_{i,j} &= \pm(u_{i\pm 1,j} - u_{i,j}), \nabla_{\pm}^y u_{i,j} = \pm(u_{i,j\pm 1} - u_{i,j}) \\ \nabla_0^x u_{i,j} &= (u_{i+1,j} - u_{i-1,j})/2, \nabla_0^y u_{i,j} = (u_{i,j+1} - u_{i,j-1})/2 \\ u_{i,j}^{n+1} &= u_{i,j}^n + \frac{\Delta t}{2\lambda} \left[ \Delta_+^x \frac{\Delta_+^x u_{ij}^n}{\sqrt{(\Delta_+^x u_{i,j}^n)^2 + (\Delta_0^y u_{i,j}^n)^2 + \tau^2}} \right. \\ &\quad \left. + \Delta_-^y \frac{\Delta_+^y u_{ij}^n}{\sqrt{(\Delta_0^x u_{i,j}^n)^2 + (\Delta_+^y u_{i,j}^n)^2 + \tau^2}} \right] - \\ &\quad \Delta t \frac{(u_0 - u^n)(\mu u_0 - \mu + 1)}{(\mu u^n - \mu + 1)^3} \end{aligned} \quad (9)$$

Among them:  $u^n$  is the musical note feature after the  $n$ th diffusion, CCC is the time interval or time step, and is a first-order, second-order recursive relationship, which is characterized in that the continuous wave signals in the upper and lower columns are both continuous. Based on this feature, good results are obtained after analyzing the fusion of partial differential equations by establishing simple equations, differential equations, and related methods.

In denoising music note features, to improve processing efficiency, it is necessary to add regularization parameters in the iterative process. Of course, in continuous iteration, the regularization parameters need to be continuously updated to achieve better results.

#### IV. MUSIC NOTE FEATURE RECOGNITION BASED ON THE HILBERT SPACE METHOD

The relational expression between each order in Hilbert space is linear, and each inter-order variable function is described and defined by itself to form a set of vector groups. Under this method, many problems can be used as time series models for solution analysis and forecasting. However, combining time series with variables in practical applications is often necessary, which is also a significant feature of the Hilbert space method [18].

Hilbert spectrum has the following characteristics:

- It can describe random processes that are linearly independent and independent of each other between two variables and have high-order convergence and stability. Therefore, the Hilbert space method can be

used to study the linear correlation between variables in a multi-period non-stationary single system.

- It can be used for analysis when the non-deterministic and unstable states are in the Hilbert space equation. The combination of non-deterministic and unstable states can be combined to build a new model to solve. Hilbert space can be used in practical observation for analysis under uncertainty and steady state instead of simply solving by differential equations.

The principle of music note feature recognition is to analyze and compare the feature vectors corresponding to different note definition domains and consecutive states. It can analyze the position relationship of different notes by calculating the matrix of words corresponding between functions and feature vectors to analyze different notes and then get different feature vectors, also known as the Hilbert space method. Based on its recognition principle, its implementation model can be obtained, as shown in Fig. 1.

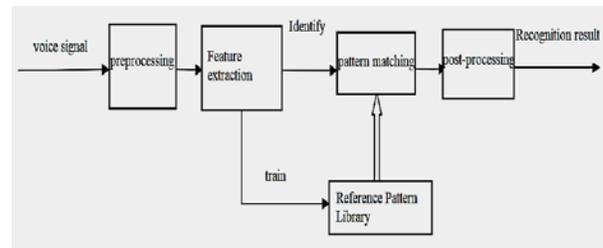


Fig. 1. Identification model

#### A. Hilbert Space Method

The Hilbert space method is a new method applicable to nonlinear and non-stationary signal processing, which can be used to study the decomposition and fitting of nonlinear problems and is of great importance in many practical applications. The Hilbert space method consists of two steps: firstly, the signal is decomposed by EMD, through which a single set of components IMF can be obtained. The IMF can effectively reflect the internal vibration pattern of the signal. Then the decomposed obtained IMF is subjected to Hilbert transform and Hilbert spectrum analysis. The results of Hilbert spectrum analysis are transformed into the corresponding state curves, and then the Hilbert space method with instantaneous time series is obtained [19].

#### B. Empirical Modal Decomposition

One of the essential concepts in the Hilbert space equation is the decomposition of modalities. In a broad sense, any new definition can be achieved using elements of different types or properties. However, for most mathematical problems, if we want to study more complex functions usually use both descriptive and approximation-type methods, but in Hilbert space equations, a special type of decomposition modality is used. The form of a continuous function is discrete into several and then a series of combinations to make it a new set of definitions [20]. EMD is the key step of the Hilbert space method, which can be regarded as the screening process to obtain IMF. The rationality and effectiveness of the EMD process are based on the following two points:

- Any complex signal  $s(t)$  can be represented as  $n$  instantaneous frequencies with actual physical meaning. In Hilbert space, each independent period can be expressed as  $n$  different frequencies, so the complex signal  $s(t)$  is represented as shown in formula (10).

$$s(t) = \sum_{k=1}^n imf_k(t) + r_n(t) \quad (10)$$

- The termination condition of the screening: Definition of IMF.

Based on the above, IMF is a signal (function) that is symmetric concerning the local zero means (Local Zero Mean). It has the same number of extreme value points (Extrema) and zero crossing points (Zero Crossings). Moreover, the graphical representation of IMF is shown in Fig. 2.

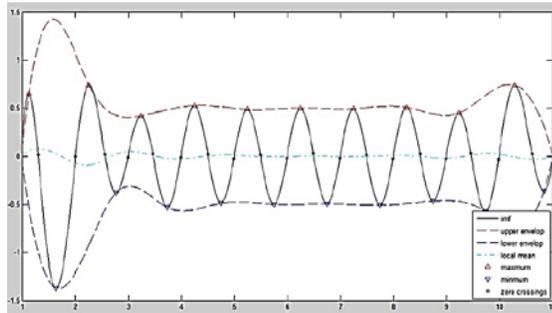


Fig. 2. IMF's diagram

### C. EMD Filtering

The Hilbert transform is superior in IMF components, and its superiority in different problems is unmatched by other modeling methods. IMF with a "single component" is integral to the Hilbert analysis process. However, most musical note signals are not IMF and do not represent Hilbert space, so it is not easy to analyze and process them when model identification is performed. Therefore, it is required that a straightforward Hilbert transform of such a signal cannot give a complete description of its frequency content. The signal has to be decomposed into IMFs so that a signal containing IMFs can be obtained [21]. The decomposition method that can effectively decompose a group of IMFs is EMD, and the essence of this decomposition method is to determine the intrinsic vibration mode of a signal based on its characteristic time scale and to define its characteristic frequency by this intrinsic vibration mode, to realize the classification of its mode. Therefore, the Hilbert transform can decompose the signal into two low-order high-order dynamic states with different Eigen frequencies and position functions. For the high-dimensional discrete spectral estimation problem, it is required to have enough time windows to ensure that the computational results converge to a suitable size range and to obtain a more accurate data set and its modal parameters (e.g., step length, step width, etc.), which is where the DMD method can have advantages in dealing with complex nonlinear models.

Based on the definition of IMF, the filtering process can be performed using the envelope formed by the signal's local maxima and local minima, respectively. Local maxima of a

signal mean decomposing a stochastic process containing all non-zero elements into a series of a small number of components and performing an exact calculation in each component to obtain an estimate of the objective function [22]. However, a definition is given for the Hilbert space: "Based on the characteristic equation (FME) algorithm proposed is a new method-Rython." The FME algorithm is an essential branch of Hilbert space, which can be used to solve signals and parameters. It has a wide range of applications in analyzing, estimating, and predicting time variation. The mean value of the upper and lower envelope of the original signal  $s(t)$  is denoted as  $m_1(t)$ . Then the difference between  $s(t)$  and  $m_1(t)$  is the first component, denoted as  $h_1(t)$ , as shown in formula (11).

$$s(t) - m_1(t) = h_1(t) \quad (11)$$

In the second screening, considering  $h_1(t)$  as the original signal and applying the same method, formula (12) is obtained.

$$h_1(t) - m_{1,1}(t) = h_{1,1}(t) \quad (12)$$

Then the screening process is similarly repeated  $k$  times until  $h_{1,k}(t)$  satisfies the conditions of IMF for the first IMF component. This process is expressed as shown in formula (13).

$$\begin{cases} h_{1,1}(t) - m_{1,2}(t) = h_{1,2}(t) \\ \vdots \\ h_{1,k-1}(t) - m_{1,k}(t) = h_{1,k}(t) \end{cases} \quad (13)$$

Let  $IMF_1(t) = h_{1,k}(t)$  so that  $IMF_1(t)$  is the first IMF component screened from the original signal  $s(t)$ , and we refer to this level of screening as inner-level screening. The inner-level screening process requires the determination of a screening termination criterion, which is an essential inner-level condition that determines whether the inner-level sequence can be extrapolated. This criterion can be bounded by the standard deviation SD (Standard Deviation) between two successive screening results. SD is defined by formula (14):

$$SD = \sum_{t=0}^T \left[ \frac{|h_{1,k-1}(t) - h_{1,k}(t)|^2}{h_{1,k-1}^2(t)} \right] \quad (14)$$

In summary, the standard EMD decomposition flow chart is shown in Fig. 3.

The Hilbert space method is a novel method applicable to nonlinear and non-smooth signal processing, which can be used to study the decomposition and fitting of nonlinear problems and is of great importance in many practical applications. The Hilbert space method consists of two steps: firstly, the signal is decomposed by EMD, through which a single component IMF can be obtained, which can effectively reflect the internal vibration pattern of the signal, and then the decomposed IMF is analyzed by Hilbert transform and Hilbert spectrum, and the result of Hilbert spectrum analysis is transformed into the corresponding state curve, and then the Hilbert space method with instantaneous time sequence [23].

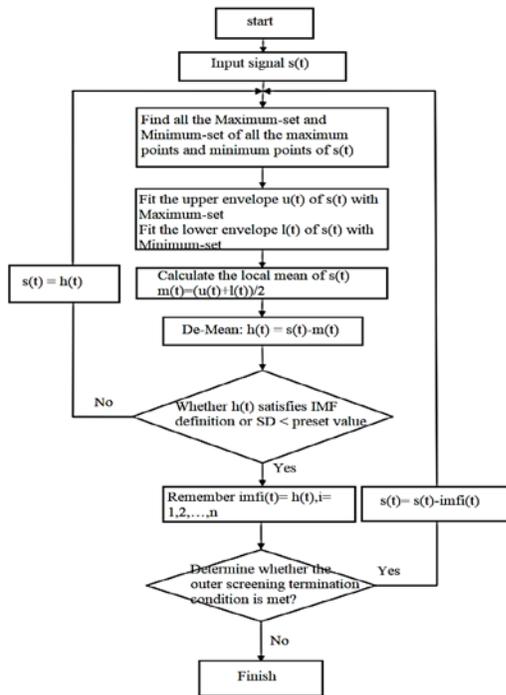


Fig. 3. EMD decomposition flow chart

After the original signal  $s(t)$  is decomposed into a set of IMFs by EMD, it can be expressed as shown in formula (15):

$$s(t) = \sum_{i=1}^n imf_i(t) + r_n(t) \quad (15)$$

This provides validity guarantees for the convenience of the Hilbert transform and the calculation of the instantaneous frequency.

The combined form of the Hilbert Marginal Spectrum (the HMS) and the instantaneous energy density level equation is the most commonly used and represent the instantaneous energy density function in the Hilbert space. The algorithm can be used to obtain the model parameters by selecting the model parameters and then transforming the complex problem into a simple mathematical one. The Hilbert marginal spectrum is a nonlinear multi-model with high prediction accuracy, and the algorithm can transform a complex problem into a linear solution. It can accurately estimate its approximate solution, as shown in formula (16). The equation of energy density level is the most commonly used and representative model parameter in Hilbert space, which can be not only directly obtained but also used for solving. It is widely used in various complex problems, and its representation is shown in the formula (17).

$$h(\omega) = \int_0^T H(\omega, t) dt \quad (16)$$

$$IE(t) = \int_{\omega} H^2(\omega, t) d\omega \quad (17)$$

The commonly used feature extraction methods are linear predictive cepstral coefficients, Mel frequency cepstral coefficients, wavelet transform-based feature extraction methods, and Hilbert space method-based feature extraction methods.

Feature extraction based on linear prediction cepstral coefficients

The feature extraction of linear predictive cepstral coefficients is the most critical and core problem in the whole multivariate statistical analysis process. We usually use single-factor models to perform dimensionality reduction in traditional regression methods. However, as the application range becomes wider and wider, the computational conditions keep improving, and the complexity of data processing becomes more and more complicated, it is challenging to meet the actual accuracy requirements, and it is impossible to directly use the variance estimation method to do multiple linear prediction verification on experimental samples. Therefore, a new linear predictive cepstral coefficient model is proposed, which can reduce the complexity of data processing by extracting feature space information with more implied parameters and higher dimensionality [24]. The feature extraction process of linear predictive cepstral coefficients is shown in Fig. 4.

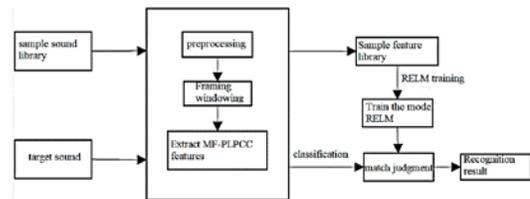


Fig. 4. Feature extraction of linear predictive cepstral coefficients

1) Feature extraction based on Mel frequency cepstrum coefficients: Mel frequency cepstrum coefficient is a nonlinear feature, and its parameters change with time. Therefore, its characteristics must be processed in some way before linear regression analysis is performed. However, the above method has some problems: firstly, it needs to calculate a large number of offline fitted curve weights and use one of them as the standard deviation, so it is computationally intensive, and the rate of change of the offline fitted curve weights will change over time. Thus it cannot accurately describe the distribution of linear feature points on a straight line. Secondly, a series of complex processes, such as obtaining new parameters, may be required before this can be used to achieve the goal of maximizing the effective extraction rate, so these are essential elements to be studied based on Mel coefficients and the feature extraction of Mel frequency cepstrum coefficients is shown in Fig. 5.

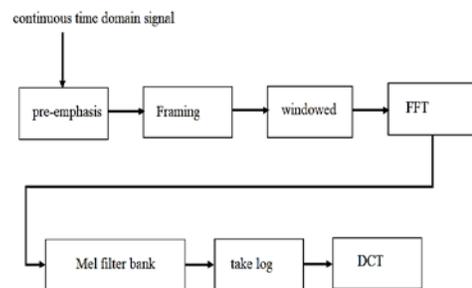


Fig. 5. Feature extraction of Mel frequency cepstral coefficients

2) *Study of wavelet transform-based feature extraction method:* As an extension of time-frequency local features in time and space, wavelet transform is a newly developed time-scale analysis method, which decomposes the image with each sub-band coefficient (i.e., distance resolution) to deal with the noise component, and the signal contains some high-frequency components. At the same time, it can effectively eliminate the low-frequency part. Moreover, wavelet transform is also a time-frequency local feature analysis method. It can extract the high-frequency components in time, space, and scale to more accurately analyze the local signal information. The feature extraction process of wavelet transform is shown in Fig. 6 [25].

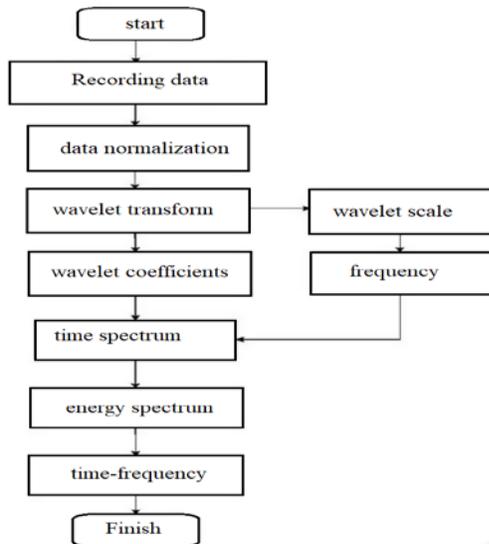


Fig. 6. Feature extraction process of wavelet transform

3) *Feature extraction method based on Hilbert space method:* The Hilbert space classification algorithm is based on the linear discriminant method of feature extraction, which is based on the traditional minimum distance classification method combined with a variety of statistical decision theories to achieve the recognition of different dimensions in the sample. The method predicts the target region by training sample data, however, due to a large amount of data and noise interference. Therefore, further improvements are needed: firstly, the original sample point information is multiplied and fitted with other discrete cosine matrices; secondly, the corresponding coordinate transformation coefficients are maximized or minimized according to the feature vectors of the classified objects, and finally, the linear discriminant method is used for parameter estimation, and the final output is combined with the Hilbert space classification method to realize the efficiency of feature extraction.

## V. CONCLUSION

To sum up, with the development of computer technology, artificial intelligence has become the trend brought by the technological progress in the new era background. With the introduction of modern new technology, the traditional method

of extracting music note features can no longer meet the needs of modern technological development. With the introduction of new technology, the traditional method of extracting musical features can no longer meet the needs of modern technology. The feature recognition method integrating partial differential equation and Hilbert space method has become the inevitable development of technology in the new era. In this paper, we propose a TV denoising model using partial differential equations, i.e., establishing its energy generalization, obtaining its Euler-Lagrange equation by variational method, and finding its numerical solution by gradient descent flow method, which can effectively protect the detailed information of music note features. In addition, when using the Hilbert space method for feature extraction and identification, the nonlinear and non-smooth feature signals can be effectively identified. It is easier and faster to use the Hilbert space method for feature identification, and the information of the feature signal can be studied accurately.

## REFERENCES

- [1] Ozan Özkan, Ali Kurt. A new method for solving fractional partial differential equations [J]. *The Journal of Analysis*,2020(2):28-30.
- [2] Paolo Marcellini. Anisotropic and p, q -nonlinear partial differential equations[J]. *Rendiconti Lincei. Scienze Fisiche e Naturali*,2020(1):31-33.
- [3] Attia Nourhane, Akgül Ali,Seba Djamila,et al. Reproducing kernel Hilbert space method for solving fractal fractional differential equations[J]. *Results in Physics*,2022(2):11-14.
- [4] Chellouf Yassamine,Maayah Banan,Momani Shaher,et al. Numerical solution of fractional differential equations with temporal two-point BVPs using reproducing kernel Hilbert space method[J]. *AIMS MATHEMATICS*,2021(4):6-9.
- [5] Hastuti Khafiizh, Syarif Arry Maulana,Fanani Ahmad Zainul,Mulyana Aton Rustandi. Natural Automatic Musical Note Player using Time-Frequency Analysis on Human Play[J]. *TELKOMNIKA (Telecommunication Computing Electronics and Control)*,2019(1):17-19.
- [6] G. Unnikrishnan. Musical Note Extraction using Self Organizing Feature Maps[J]. *International Journal of Computer Applications*,2018(9):182-184.
- [7] Heihoff Frederic. Generalized solutions for a system of partial differential equations arising from urban crime modeling with a logistic source term[J]. *Zeitschrift für angewandte Mathematik und Physik*,2020(3):71-77.
- [8] Xia Miao, Ziyao Yu, Ming Liu. "Using Partial Differential Equation Face Recognition Model to Evaluate Students' Attention in a College Chinese Classroom", *Advances in Mathematical Physics*, vol. 2021, Article ID 3950445, 10 pages, 2021.
- [9] Tamboli, Allabakash Isak and Kokate, Rajendra D.. "An Effective Optimization-Based Neural Network for Musical Note Recognition" *Journal of Intelligent Systems*, vol. 28, no. 1, 2019, pp. 173-183.
- [10] Arvind Kumar et al., Hilbert Spectrum Based Features for Speech/Music Classification, *SERBIAN JOURNAL OF ELECTRICAL ENGINEERING* Vol. 19, No. 2, June 2022, 239-259
- [11] Qin Lin, "Music Score Recognition Method Based on Deep Learning", *Computational Intelligence and Neuroscience*, vol. 2022, Article ID 3022767, 12 pages, 2022.
- [12] Alías F, Socoró JC, Sevillano X. A Review of Physical and Perceptual Feature Extraction Techniques for Speech, Music and Environmental Sounds. *Applied Sciences*. 2016; 6(5):143.
- [13] Li, X., Zheng, Y. & Li, X. Extraction of novel features for emotion recognition. *J. Shanghai Univ.(Engl. Ed.)* **15**, 479 (2011).
- [14] Pingli Sun, Chenxia Wang, Min Li, Lanqi Liu, "Partial Differential Equations-Based Iterative Denoising Algorithm for Movie Images", *Advances in Mathematical Physics*, vol. 2021, Article ID 8176746, 10 pages, 2021. <https://doi.org/10.1155/2021/8176746>

- [15] Yvain Bruned, Martin Hairer, Lorenzo Zambotti. Renormalisation of Stochastic Partial Differential Equations[J]. EMS Newsletter, 2020:115-118.
- [16] Konstantinos Dareiotis, Benjamin Gess. Supremum estimates for degenerate, quasilinear stochastic partial differential equations[J]. Annales de l'Institut Henri Poincaré, Probabilités et Statistiques, 2019(3):55-57.
- [17] Mohammadreza Foroutan, Raheleh Asadi, Ali Ebadian. A reproducing kernel Hilbert space method for solving the nonlinear three-point boundary value problems[J]. International Journal of Numerical Modelling: Electronic Networks, Devices and Fields, 2019(3):32-34.
- [18] Asia Khalaf Albzeirat, Muhammad Zaini Ahmad, Shaher Momani, Israr Ahmad. New implementation of reproducing kernel Hilbert space method for solving a fuzzy integro-differential equation of integer and fractional orders[J]. Journal of King Saud University - Science, 2018(3):30.
- [19] Omar Abu Arqub, Banan Maayah. Modulation of reproducing kernel Hilbert space method for numerical solutions of Riccati and Bernoulli equations in the Atangana-Baleanu fractional sense[J]. Chaos, Solitons and Fractals: the interdisciplinary journal of Nonlinear Science, and Nonequilibrium and Complex Phenomena, 2019(02):125.
- [20] Seda Cayan, Mehmet Sezer. PELL POLYNOMIAL APPROACH FOR DIRICHLET PROBLEM RELATED TO PARTIAL DIFFERENTIAL EQUATIONS[J]. Journal of Science and Arts, 2019(3):19-22.
- [21] Ricardo Nochetto, Stefan Sauter, Christian Wieters. Space-time Methods for Time-dependent Partial Differential Equations[J]. Oberwolfach Reports, 2018(1):14-16.
- [22] Akgül, Esra Karatas. Solutions of Nonlinear System of Differential Equations by Reproducing Kernel Hilbert Space Method[J]. Journal of Advanced Physics, 2018(1):7-10.
- [23] Younis A.N., Ramo F.M.. A new parallel bat algorithm for musical note recognition[J]. International Journal of Electrical and Computer Engineering, 2021(1):11-15.
- [24] Abhijit V. Chitre, Aditya Abhyankar. Mutual Information Based Framework of Quality Parameter Formulation for Similarity Index between Two Musical Notes[J]. International Journal of Recent Technology and Engineering (IJRTE), 2020(5):8-15.
- [25] Mike Pacheco, Heather Glynn Crawford-Ferre, Henry King. math by the month: Musical notes[J]. Teaching Children Mathematics, 2018(6):24-26.

# Hyperparameter Optimization of Support Vector Regression Algorithm using Metaheuristic Algorithm for Student Performance Prediction

M. Riki Apriyadi<sup>1</sup>, Ermatita<sup>2\*</sup>, Dian Palupi Rini<sup>3</sup>

Doctoral Student of Engineering Science, Faculty of Engineering  
Universitas Sriwijaya, Palembang, Indonesia<sup>1</sup>

Faculty of Computer Science, Universitas Sriwijaya, Palembang, Indonesia<sup>2,3</sup>

**Abstract**—Improving student learning performance requires proper preparation and strategy so that it has an impact on improving the quality of education. One of the preparatory steps is to make a prediction modeling of student performance. Accurate student performance prediction models are needed to help teachers develop the potential of diverse students. This research aims to create a predictive model of student performance with hyperparameter optimization in the Support Vector Regression Algorithm. The hyperparameter optimization method used is the Metaheuristic Algorithm. The Metaheuristic Algorithms used in this study are Particle Swarm Optimization (PSO) and Genetic Algorithm (GA). After obtaining the best SVR hyperparameter, the next step is to model student performance predictions, which in this study produced two models, namely PSVR Modeling and GSVR Modeling. The resulting predictive modeling will also be compared with previous researchers' prediction modeling of student performance using five models: Support Vector Regression, Naïve Bayes, Neural Network, Decision Tree, and Random Forest. The regression performance metric parameter, Root Mean Square Error (RMSE), evaluates modeling results. The test results show that predictive student performance using PSVR Modeling produces the smallest RMSE value of 1.608 compared to predictions of student performance by previous researchers so that the proposed prediction model can be used to predict student performance in the future.

**Keywords**—Student performance; feature selection; particle swarm optimization; genetic algorithm; support vector regression

## I. INTRODUCTION

Educators need a prediction of student performance to improve student achievement. Predicting student performance is used as material for evaluating student learning so that it can facilitate the diversity of potential students, both those who excel academically [1][2] and detect students who have the potential to experience failure [3]. Accurate prediction of student performance can also be the right policy decision in educational institutions [4].

The implementation of the Machine Learning Algorithm to predict student performance is to compare the accuracy of both classification and regression [5]. The Machine Learning Algorithms used include Neural Networks, Decision Trees, Naïve Bayes, SVM, KNN, and Logistic Regression [2][5][6]. Support Vector Machine (SVM) is a Machine Learning Algorithm that can be used to predict student performance

[5][7][8][9]. As for solving regression problems, SVM is better known as Support Vector Regression (SVR) [10]. SVR has good generalization ability, can be implemented for non-linear data with high dimensions, and has low computational complexity [11]. In addition, other advantages of SVR are overcoming overfitting and making predictions with data that is not too large [12]. From these advantages, SVR can be implemented in this study to predict student performance [11]. Problems often experienced by SVR occur in large-scale data, thus making significant computational processes challenging to determine optimal hyperparameter values [11][13].

Optimal selection of hyperparameters in Machine Learning Algorithms has been carried out using various Metaheuristic Algorithm approaches, namely Particle Swarm Optimization (PSO)[14], Artificial Bee Colony (ABC) [15], and Genetic Algorithm (GA) [16]. The SVR Algorithm is a Machine Learning Algorithm; optimizing hyperparameters in SVR modeling will increase the value of modeling accuracy [16][17].

## II. RELATED WORK

Many researchers have researched student performance prediction using Machine Learning Algorithms. Tomasevic et al. [6] predict student performance by comparing Machine Learning Algorithm modeling, namely KNN, SVM, ANN, Decision tree, Naïve Bayes, and Logistic Regression with classification and regression models. This study used data on students' past learning achievements, learning engagement, search activities, discussion participation, and demographics. The results of this study show that ANN outperforms other Machine Learning Algorithms with the best accuracy.

In the study of student performance prediction conducted by Xu et al. [18], using student activity data, as many as 4,000 students on the online duration, traffic volume, and connection frequency. The resulting classification prediction modeling is in the form of passed and failed. The Machine Learning Algorithms used include Decision Trees, ANN, and SVM. This study showed that the ANN and SVM Algorithms for predicting student achievement were the most accurate.

Cortez et al. [5] compared the accuracy of predicting student performance with classification and regression modeling using Neural Network, Decision Tree, Naïve Bayes, Random Forest, and SVM Algorithms to predict student

\*Corresponding Author.

performance in mathematics and Portuguese. The experimental results show that in the classification case, the Naïve Bayes Algorithm produces the best accuracy for predicting student performance in mathematics and the Decision Tree Algorithm produces the best accuracy for predicting student performance in Portuguese. In the regression case, the Random Forest Algorithm has the best accuracy for predicting student performance in mathematics. In contrast, the Naïve Bayes and Random Forest Algorithms produce the best accuracy for predicting student performance in Portuguese.

This study will use previous research datasets, namely the performance of high school students in Portugal in mathematics [5], with the development of the SVR Algorithm. The choice of the SVR algorithm is because the algorithm can overcome overfitting and make predictions with data that is not too large [12]. The development of the SVR Algorithm is to find optimal hyperparameters in the SVR Algorithm using the Metaheuristic Algorithm, namely PSO and GA [14][16]. By using optimal hyperparameters, the application of the SVR Algorithm can increase the accuracy of predictive modeling [16][17]. So the proposed contribution of this research is developing a model predicting student performance on the SVR Algorithm with hyperparameter optimization using the Metaheuristic Algorithm, which previous researchers have not done.

### III. MATERIAL AND METHOD

In this study, there are several stages needed to predict student performance. In the early stages, the collection of the student performance dataset, Split dataset to data training and data testing, Optimization of hyperparameters on SVR using the Metaheuristic Algorithm, and Modeling of Student Performance Prediction.

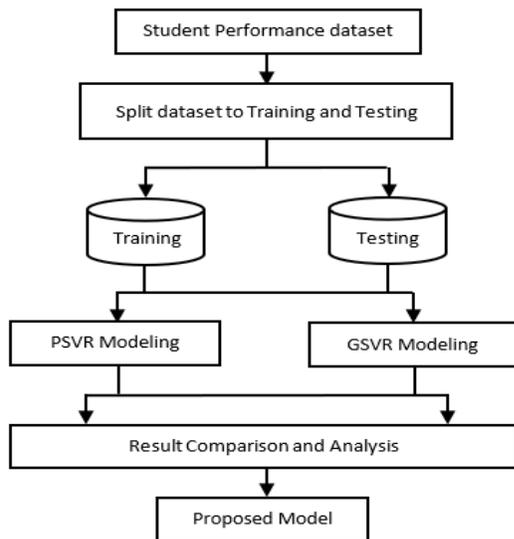


Fig. 1. The proposed method

In Fig. 1, it can be seen the stages carried out in this study. The first step was collecting the dataset. Dataset collection can be done by downloading from the UCI Machine Learning Website<sup>1</sup>. After processing the dataset, the dataset will be split into training and testing data, with 90% of the training data and

10% of the testing data. Training data is used when training Algorithms and looking for suitable models, while data testing is used as test data to determine the performance of the model that has been produced. The next step in this study is to model student performance predictions using the SVR Algorithm with hyperparameter optimization. This study used two models: GSVR Modeling and PSVR Modeling. After predictive modeling is generated, the next step is to evaluate the performance of the regression performance metrics using RMSE and compare it with Machine Learning Algorithm modeling done by previous researchers using the same dataset [5].

#### A. Data for Student Performance Prediction

This study uses a dataset of student performance at secondary schools in Portugal from the UCI Machine Learning Repository Website. This collection of student performance data comes from two secondary schools, namely the Gabriel Pereira School and the Mousinho da Silveira School mathematics. It consists of 395 instances and 33 demographic, social, financial, and academic data attributes [5]. Of the 33 attributes in this dataset, one attribute is the result of students' mathematics final exam scores, namely G3, which will be used as the target for modeling student performance predictions, so 32 attributes in the dataset affect student performance. The description of the student performance dataset used can be seen in Table I.

TABLE I. FEATURES AND DESCRIPTION OF THE DATASET

Feature	Description
school	student's school
sex	student's sex
age	student's age
address	student's home address type
famsize	family size
Pstatus	parent's cohabitation status
Medu	mother's education
Fedu	father's education
Mjob	mother's job
Fjob	father's job
reason	reason to choose this school
guardian	student's guardian
traveltime	home to school travel time
studytime	weekly study time
failures	number of past class failures
schoolsup	extra educational support
famsup	family educational support
paid	extra paid classes within the course subject
activities	extra-curricular activities
nursery	attended nursery school

<sup>1</sup> [https://archive.ics.uci.edu/ml/Datasets/student\\_performance](https://archive.ics.uci.edu/ml/Datasets/student_performance)

higher	wants to take higher education
internet	Internet access at home
romantic	With a romantic relationship
famrel	Quality of family relationships
freetime	free time after school
goout	going out with friends
Dalc	workday alcohol consumption
Walc	weekend alcohol consumption
health	current health status
absences	number of school absences
G1	first-period grade
G2	second-period grade
G3	final grade (target of prediction)

### B. Machine Learning Algorithms

This research will focus on developing the SVR Algorithm by optimizing hyperparameters using the Metaheuristic Algorithm. At the end of the development, we will compare with other Machine Learning Algorithms used by previous researchers to predict student performance, namely Support Vector Regression (SVR), Naïve Bayes, Neural Networks, Decision Trees, and Random Forests [5].

1) *Support vector regression (SVR)*: SVR is a development of the SVM Algorithm introduced by Vladimir Naumovich Vapnik in 1995 [19]. SVR shows good performance in solving regression problems [11]. SVR applies the Structural Risk Minimization (SRM) method, which is a method with a focus on finding the optimal hyperplane and minimizing errors from the training data and incentive loss function, resulting in a continuous and real-value data output [20]. In this study, the hyperparameters used are C, gamma, and epsilon.

2) *Naïve Bayes*: Naïve Bayes is a simple probabilistic classifier that calculates a set of probabilities by summing the frequencies and combinations of values from the given dataset [21]. This Algorithm uses Bayes theorem and assumes that all attributes are independent or not interdependent, given the value of the class variable [22]. Naive Bayes is based on the simplifying assumption that attribute values are conditionally independent when given output values [23]. In other words, given the output values, the probabilities of observing together are the product of the individual probabilities [24].

3) *Neural networks*: Neural networks are information processing Algorithms inspired by the workings of the biological nervous system, especially in human brain cells, in processing information [25]. Neural networks consist of many information-processing elements that are connected and work together to solve a particular problem, which is generally a classification or prediction problem [26].

4) *Decision tree*: A Decision tree is a predictive model technique used for task classification and prediction [27]. A

Decision tree divides the problem search space into problems [28]. The process in the decision tree is to change the form of table data into a tree model. The model tree will generate rules and be simplified [29].

5) *Random forest*: Random Forest Random Forest is a supervised learning Algorithm released by Breiman [30]. Random Forest is commonly used to solve problems related to classification, regression, etc. This Algorithm is a combination of several tree predictors, or it can be called a decision tree, where each tree depends on a random vector value sampled freely and evenly on all trees in the forest [31]. The prediction results from the Random Forest get the most results from each decision tree [32].

### C. Metaheuristic Algorithm for Optimizing Hyperparameters in the SVR Algorithm

A Metaheuristic can be defined as an iterative generation process that guides subordinate heuristics by intelligently combining different concepts to exploit the search space used to organize information in efficiently finding a near-optimal solution [33]. A Metaheuristic Algorithm is used to help optimally find hyperparameters to produce the best accuracy value in predictive modeling [12][17]. Determining hyperparameters in a Machine Learning Algorithm is a significant step in modeling [16]. The optimal hyperparameter is determined based on the fitness function. The fitness function is as follows.

$$\text{Fitness function} = \sqrt{\sum_{i=1}^n (\hat{y}_i - y_i)^2} \quad (1)$$

Where  $\hat{y}_i$  is the predicted value,  $y_i$  is the original value of the sample dataset, and  $n$  is the total number of samples. In this study, The Metaheuristic Algorithms used to find optimal Hyperparameters in SVR are PSO and GA, which will be discussed as follows.

1) *Particle swarm optimization (PSO)*: PSO has been developed by Kennedy and Eberhart as an optimization Algorithm [34]. The way PSO works is based on the results of observing the social behavior of a group of birds and fish moving to a specific position to get food, which is then referred to as the best position in the multidimensional search space [35][36]. The term particle denotes a bird in a flock that collectively influences its intelligence or that of the group [37]. According to the search area, particle movement with velocity will save it as the best position as Pbest and Gbest [38]. PSO aims to get the optimal solution by minimizing the fitness function [39].

In this study, we will apply PSO as an optimizer for SVR hyperparameters, with the name PSVR Modeling. The steps taken are the initialization of the initial parameters of the PSO in the form of particle velocity, initial particle position, and iteration. Particles will update the position and velocity memory to obtain the Pbest and Gbest values [11]. The best fitness value of the iteration limit will produce the best SVR hyperparameter combination in the form of C, gamma, and epsilon in PSVR Modeling.

2) *Genetic algorithm (GA)*: GA is an evolutionary Algorithm inspired by the mechanism of natural selection based on Charles Darwin's theory [40]. GA was introduced in 1975 at the University of Michigan by John Holland [41]. GA is widely used to solve optimization problems [42]. GA works to find the optimum solution simultaneously at several points in one generation, and then GA manipulates the population structure symbolically as the best solution [43]. In GA, a solution is a chromosome, and a group of chromosomes is called a population. Chromosomes from one population form a new population based on the objective function or the best fitness value [44].

This study will also apply GA as an optimizer for SVR hyperparameters, with the name GSVR Modeling. The steps are to initialize the initial GA parameters in the form of the initial population and iteration limits. The initial population in the state of individuals will be selected based on the order of the best fitness function with the selection stages [16]. After that, the cross-over stage is carried out, namely the exchange of genes between one chromosome and another based on the parameter of the crossover rate. The next stage is a mutation, in which the resulting chromosome will replace one or more genes with other genes at random [45]. In the final stage, a new individual will be generated to determine the best fitness value obtained from the iteration limit to produce the best SVR hyperparameter combination in the form of C, gamma, and epsilon in GSVR Modeling.

#### D. Evaluation Method

The developed model will be evaluated using regression performance metric parameters in the final stage. The function of the regression performance metric parameter is to measure the accuracy of modeling predictions of student performance. This study's regression performance metric parameter is the Root Mean Square Error (RMSE). RMSE can be defined as the square root of the average value of squared errors between the actual value and the forecast value [36].

$$RMSE = \sqrt{\frac{1}{n} \sum_{j=1}^n (y_j - \hat{y}_j)^2} \quad (2)$$

Where  $\hat{y}_j$  is the predicted value of student performance,  $y_j$  is the original value of the student performance sample dataset, and  $n$  is the total number of samples.

#### IV. RESULT AND DISCUSSION

This study uses the student performance dataset by Cortez [5], focusing on students' final exam scores in Mathematics. This student performance dataset from two secondary schools consists of 395 instances and 33 attributes. This dataset is a file with the Comma Separated Values (CSV) format in excel. After checking each row and column of data, no empty data is found in this dataset, so it can be stated that this dataset has been filled in completely. The next step is to model student performance predictions by optimizing hyperparameters in SVR Algorithm using the Metaheuristic Algorithm, which is then compared with modeling student performance predictions that have been carried out by previous researchers using Machine Learning Algorithms.

#### A. Results of Previous Research Results using Machine Learning Algorithms

Cortez et al. [5] have researched predicting student performance modeling using Machine Learning Algorithms, including SVR, Naïve Bayes, Neural Networks, Decision Trees, and Random Forests. The modeling results that have been produced are in Table II below.

TABLE II. RESULT OF PREVIOUS RESEARCH [5]

Algorithm	RMSE
SVR	2.09
Naïve Bayes	2.01
Neural Networks	2.05
Decision Tree	1.94
Random Forest	1.75

In Table II, the modeling results show that the best predictive modeling of student performance is obtained in modeling predictive student performance in the Random Forest Algorithm with an RMSE value of 1.75, while the worst student performance in the SVR Algorithm with an RMSE value of 2.09. In previous studies, only making comparisons of the accuracy of modeling predictions of student performance using Machine Learning Algorithms and hyperparameter optimization has not been carried out on Machine Learning Algorithms. So that in this study will improve the accuracy of modeling student performance predictions with the SVR Algorithm by optimizing hyperparameters using the Metaheuristic Algorithm.

#### B. Results of Optimization Hyperparameter SVR using Metaheuristic Algorithm and Modeling

This stage is the initial stage for developing the SVR Algorithm to predict student performance. Hyperparameter optimization is performed to determine the best hyperparameter composition of the SVR Algorithm as a predictive model to be developed. The settings for the hyperparameter values to be optimized are C, gamma, and epsilon by determining the range of upper and lower limit values, with hyperparameter values C = [100 - 1000], gamma = [0.001 - 0.009], and epsilon = [0.001 - 0.009]. The predictive modeling of student performance resulting from hyperparameter optimization using the Metaheuristic Algorithm is as follows.

1) *Optimization hyperparameter SVR using particle swarm optimization (PSVR modeling)*: In this study, PSO will be applied as an optimizer for SVR hyperparameters. The SVR hyperparameters are C, gamma, and epsilon. Meanwhile, as an optimization Algorithm, the PSO parameters will be determined by initialization, also carried out on the initial PSO parameters. The PSO parameter will be set with a total of 50 particles, while the value of C1 is 1.0 and C2 is 2.0, with a weight value of W is 0.5. The number of iterations in this research will be varied by the number of iterations of 50, 100, 250, and 500.

TABLE III. RESULT OF PSVR MODELING

Iteration	Optimal Hyperparameter			RMSE
	C	gamma	epsilon	
50	380	0.001	0.006	1.681
100	103	0.002	0.001	1.608
250	358	0.001	0.008	1.683
500	450	0.001	0.005	1.675

Based on the PSVR Modeling results in Table III, the optimal SVR hyperparameter combination was obtained at C, gamma, epsilon = [103, 0.002, 0.001] at the 100th iteration is the best RMSE value of 1.608. The selection of optimal hyperparameters generated in PSVR Modeling results from searching for a combination of hyperparameter limits on SVR that has been determined using the PSO method search stages. In the PSO method, the resulting hyperparameter combination will be evaluated for its fitness value based on the Pbest and Gbest values of the iterations and predetermined PSO parameters so that the best hyperparameter combination with the smallest RMSE value is obtained.

2) Optimization hyperparameter SVR using genetic algorithm (GSVR modeling): GA will be applied as an optimizer for SVR hyperparameters in this study. The SVR hyperparameters are C, gamma, and epsilon. Meanwhile, as an optimization Algorithm, the GA parameters will be determined by initialization, also carried out on the initial GA parameters. The GA parameter will be determined by a total of 50 individuals, while the mutation coefficient value is 0.01, with a cross-over coefficient value of 0.5. The number of iterations in this research will be varied by the number of iterations of 50, 100, 250, and 500.

TABLE IV. RESULT OF GSVR MODELING

Iteration	Optimal Hyperparameter			RMSE
	C	gamma	epsilon	
50	102	0.001	0.006	1.831
100	101	0.001	0.008	1.831
250	100	0.001	0.008	1.830
500	101	0.001	0.007	1.831

Based on the GSVR Modeling results in Table IV, the best RMSE value was 1.830 at the 250th iteration with the optimal SVR hyperparameter combination at C, gamma, epsilon = [100, 0.001, 0.008]. In GSVR Modeling, selecting optimal hyperparameters is also the result of searching for a combination of hyperparameter limits on a predetermined SVR using the GA method search stages. In the GA method, the resulting hyperparameter combination will be evaluated for its fitness value based on the GA stages to get the best individual based on the iteration results and GA cycle stages in the form of selection, cross-over, and mutation so that the best hyperparameter combination with the smallest RMSE value is obtained.

From the experimental results, obtained modeling of student performance with PSVR Modeling and GSVR Modeling, which will then be used to be compared with student performance modeling in previous studies.

### C. Comparing and Analysis Results

In this study, modeling of student performance predictions has been carried out using the SVR Algorithm, which is optimized for hyperparameters with the Metaheuristic Algorithm, by producing a proposed model in the form of PSVR Modeling in Table III and GSVR Modeling in Table IV. In the PSVR Modeling experiment with the optimal SVR hyperparameter, the RMSE value was 1.608, and GSVR Modeling with optimal SVR hyperparameters produced an RMSE value of 1.830, so when compared to modeling student performance predictions using the Machine Learning Algorithm obtained from previous research in Table II, which will produce a comparison like in Fig. 2.

Fig. 2 shows that modeling predictive student performance with PSVR Modeling gets the best results with the smallest RMSE value of 1.608 compared to the RMSE value on GSVR modeling and performance prediction modeling students conducted in previous studies. It can also be seen that the SVR algorithm gets the highest RMSE value of 2.09 compared to the Machine Learning algorithm used in previous studies for modeling student performance predictions. With this research, it can be seen that optimizing the hyperparameters in the SVR algorithm can reduce the error value or increase the accuracy of modeling student performance predictions.

Fig. 3 shows the graph of increasing accuracy using hyperparameter optimization using the Metaheuristic Algorithm for the SVR Algorithm. The experimental results show an increase in the accuracy of the SVR algorithm with the proposed model. GSVR Modeling shows an increase in accuracy of 12.44% with a decrease in the error value with RMSE from 2.09 to 1.830 compared to the SVR Algorithm. As a comparison, the best improvement is PSVR Modeling which shows an increase in accuracy of 23.06% with a decrease in the error value with RMSE from 2.09 to 1.608 compared to the SVR Algorithm.

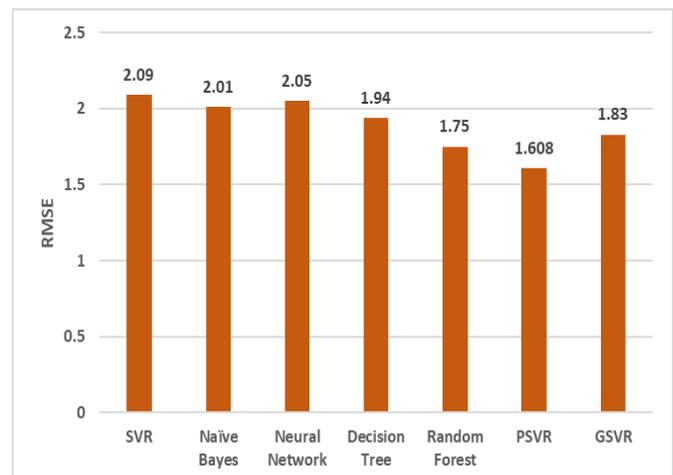


Fig. 2. Comparison of the results modeling student performance prediction

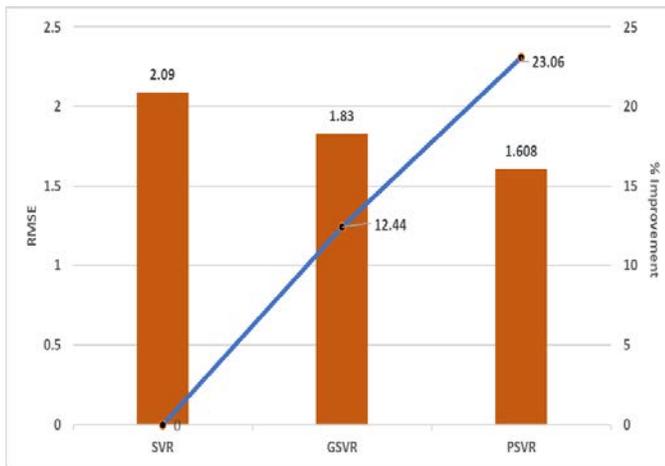


Fig. 3. Percentage Improvement of SVR with hyperparameter optimization using metaheuristic algorithm

## V. CONCLUSION

Modeling student performance predictions using an accurate Machine Learning Algorithm can predict student performance so that appropriate strategies can be determined to improve student learning outcomes. In previous research [5], modeling predictions of student performance were compared using several machine learning algorithms. This study has developed a predictive model for student performance by optimizing the hyperparameters in SVR using the Metaheuristic Algorithm, namely PSO and GA, to produce a proposed model with two models, PSVR Modeling and GSVR Modeling. In predicting student performance predictions using PSVR Modeling, the prediction accuracy is the best compared to predicting student performance using other Machine Learning Algorithms with an RMSE value of 1.608. The increase in the accuracy of the RMSE value was also generated by modeling predictions of student performance with PSVR Modeling of 23.06% compared to predictions by modeling student performance using the SVR Algorithm. This experiment shows that the student performance prediction model with the proposed model can be used to predict student performance in the future. In this study, the selection of optimal hyperparameters in the SVR Algorithm has been proven to increase accuracy in predicting student performance. Future research is expected to be able to conduct experiments by setting hyperparameters on C, gamma, and epsilon with a more varied range of values so that it is possible to obtain even better predictive modeling accuracy results.

## VI. FUTURE WORK

In future research, further the development will be carried out on predicting student performance modeling using the Feature Selection Method with Metaheuristic Algorithms. So modeling student performance predictions using the feature selection method will produce features that influence student performance predictions and increase the accuracy of the resulting model.

## ACKNOWLEDGMENT

The first author is a doctoral student at the Faculty of Engineering, Universitas Sriwijaya. The authors would like to

thank Universitas Sriwijaya for their support in carrying out this research.

## REFERENCES

- [1] C. F. Rodríguez-Hernández, M. Musso, E. Kyndt, and E. Cascallar, "Artificial neural networks in academic performance prediction: Systematic implementation and predictor evaluation," *Comput. Educ. Artif. Intell.*, vol. 2, no. December 2020, 2021, doi: 10.1016/j.caeai.2021.100018.
- [2] H. A. Mengash, "Using data mining techniques to predict student performance to support decision making in university admission systems," *IEEE Access*, vol. 8, pp. 55462–55470, 2020, doi: 10.1109/ACCESS.2020.2981905.
- [3] S. Helal et al., "Predicting academic performance by considering student heterogeneity," *Knowledge-Based Syst.*, vol. 161, pp. 134–146, 2018, doi: 10.1016/j.knosys.2018.07.042.
- [4] H. Waheed, S. U. Hassan, N. R. Aljohani, J. Hardman, S. Alelyani, and R. Nawaz, "Predicting academic performance of students from VLE big data using deep learning models," *Comput. Human Behav.*, vol. 104, 2019, doi: 10.1016/j.chb.2019.106189.
- [5] P. Cortez and A. Silva, "Using data mining to predict secondary school student performance," *15th Eur. Concurr. Eng. Conf. 2008, ECEC 2008 - 5th Futur. Bus. Technol. Conf. FUBUTEC 2008*, vol. 2003, no. 2000, pp. 5–12, 2008.
- [6] N. Tomasevic, N. Gvozdenovic, and S. Vranes, "An overview and comparison of supervised data mining techniques for student exam performance prediction," *Comput. Educ.*, vol. 143, p. 103676, 2019, doi: 10.1016/j.compedu.2019.103676.
- [7] A. Charleonnann, T. Fufaung, T. Niyomwong, W. Chokchueypattanakit, S. Suwannawach, and N. Ninchawee, "Predictive analytics for chronic kidney disease using machine learning techniques," *Manag. Innov. Technol. Int. Conf., pp. MIT80–MIT83*, 2016, doi: 10.1109/MITICON.2016.8025242.
- [8] S. Oloruntoba and J. Akinode, "Student Academic Performance Prediction Using Support Vector Machine," *Int. J. Eng. Sci. Res. Technol.*, vol. 6, no. 12, pp. 588–598, 2017.
- [9] I. Burman and S. Som, "Predicting Students Academic Performance Using Support Vector Machine," *Amity Int. Conf. Artif. Intell.*, pp. 756–759, 2019.
- [10] J. H. Fu, J. H. Chang, Y. M. Huang, and H. C. Chao, "A support vector regression-based prediction of students' school performance," *Proc. - 2012 Int. Symp. Comput. Consum. Control. IS3C 2012*, pp. 84–87, 2012, doi: 10.1109/IS3C.2012.31.
- [11] H. Xu, "Prediction of Students' Performance Based on the Hybrid IDA-SVR Model," *Complexity*, vol. 2022, 2022, doi: 10.1155/2022/1845571.
- [12] H. Harafani, "Parameter Optimization of Support Vector Machine Based on Genetic Algorithm for Forest Fire Estimation," *Journal of Intelligent Systems.*, vol. 1, no. 2, pp. 83–90, 2015.
- [13] S. R. N and P. C. Deka, "Support vector machine applications in the field of hydrology: A review," *Appl. Soft Comput. J.*, vol. 19, pp. 372–386, 2014, doi: 10.1016/j.asoc.2014.02.002.
- [14] X. Wang, J. Wen, Y. Zhang, and Y. Wang, "Real estate price forecasting based on SVM optimized by PSO," *Optik (Stuttg.)*, vol. 125, no. 3, pp. 1439–1443, 2014, doi: 10.1016/j.ijleo.2013.09.017.
- [15] A. Kumar, G. Kabra, E. K. Mussada, M. K. Dash, and P. S. Rana, "Combined artificial bee colony algorithm and machine learning techniques for prediction of online consumer repurchase intention," *Neural Comput. Appl.*, vol. 31, pp. 877–890, 2017, doi: 10.1007/s00521-017-3047-z.
- [16] A. S. Wicaksono and A. A. Supianto, "Hyperparameter optimization using genetic algorithm on machine learning methods for online news popularity prediction," *Int. J. Adv. Comput. Sci. Appl.*, vol. 9, no. 12, pp. 263–267, 2018, doi: 10.14569/IJACSA.2018.091238.
- [17] Z. Luo, M. Hasanippanah, H. Bakhshandeh Amnieh, K. Brindhadevi, and M. M. Tahir, "GA-SVR: a novel hybrid data-driven model to simulate vertical load capacity of driven piles," *Eng. Comput.*, vol. 37, no. 2, pp. 823–831, 2019, doi: 10.1007/s00366-019-00858-2.
- [18] X. Xu, J. Wang, H. Peng, and R. Wu, "Prediction of academic

- performance associated with internet usage behaviors using machine learning algorithms,” *Comput. Human Behav.*, vol. 98, no. January, pp. 166–173, 2019, doi: 10.1016/j.chb.2019.04.015.
- [19] C. Cortes and V. Vapnik, “Support-Vector Networks,” vol. 20, pp. 273–297, 1995, doi: 10.1109/64.163674.
- [20] A. J. Smola and B. Scholkopf, “A tutorial on support vector regression,” *Stat. Comput.*, vol. 14, pp. 199–222, 2004, doi: 10.1210/me.10.7.813.
- [21] U. N. Dulhare, “Prediction system for heart disease using Naive Bayes and particle swarm optimization,” *Biomed. Res.*, vol. 29, no. 12, pp. 2646–2649, 2018, doi: 10.4066/biomedicalresearch.29-18-620.
- [22] A. Yasar and M. M. Saritas, “Performance Analysis of ANN and Naive Bayes Classification Algorithm for Data Classification,” *Int. J. Intell. Syst. Appl. Eng.*, vol. 7, no. 2, pp. 88–91, 2019, doi: 10.18201/ijisae.2019252786.
- [23] D. Kabakchieva, “Predicting student performance by using data mining methods for classification,” *Cybern. Inf. Technol.*, vol. 13, no. 1, pp. 61–72, 2013, doi: 10.2478/cait-2013-0006.
- [24] S. L. Ting, W. H. Ip, and A. H. C. Tsang, “Is Naïve bayes a good classifier for document classification?,” *Int. J. Softw. Eng. its Appl.*, vol. 5, no. 3, pp. 37–46, 2011.
- [25] J. Feng, “Predicting Students’ Academic Performance with Decision Tree and Neural Network,” 2019, [Online]. Available: <https://www.semanticscholar.org/paper/11296c25a376b1ef8e3520460d8fac5f09a4d1fc>.
- [26] P. B. Kumar and V. Boddu, “Analysis Of The Impact Of Social And Emotional Factors On Student Performance Using Data Mining Techniques,” 2018, [Online]. Available: <https://www.semanticscholar.org/paper/dbd8ed80435374647aabe4600d7a2869dd54e640>.
- [27] H. Hamsa, S. Indiradevi, and J. J. Kizhakkethottam, “Student Academic Performance Prediction Model Using Decision Tree and Fuzzy Genetic Algorithm,” *Procedia Technol.*, vol. 25, pp. 326–332, 2016, doi: 10.1016/j.protcy.2016.08.114.
- [28] H. Rao et al., “Feature selection based on artificial bee colony and gradient boosting decision tree,” *Appl. Soft Comput. J.*, 2018, doi: 10.1016/j.asoc.2018.10.036.
- [29] M. Hussain, W. Zhu, W. Zhang, and S. M. R. Abidi, “Student Engagement Predictions in an e-Learning System and Their Impact on Student Course Assessment Scores,” *Comput. Intell. Neurosci.*, vol. 2018, 2018, doi: 10.1155/2018/6347186.
- [30] S. Rebai, F. Ben Yahia, and H. Essid, “A graphically based machine learning approach to predict secondary schools performance in Tunisia,” *Socioecon. Plann. Sci.*, vol. 70, no. August 2018, p. 100724, 2020, doi: 10.1016/j.seps.2019.06.009.
- [31] R. Ghorbani and R. Ghousi, “Comparing Different Resampling Methods in Predicting Students’ Performance Using Machine Learning Techniques,” *IEEE Access*, vol. 8, pp. 67899–67911, 2020, doi: 10.1109/ACCESS.2020.2986809.
- [32] E. Osmanbegović, M. Suljic, and H. Agić, “Determining Dominant Factor For Students Performance Prediction By Using Data Mining Classification Algorithms,” 2014, [Online]. Available: <https://www.semanticscholar.org/paper/2311afce19ecd511ca7d92dd12ac77f79fe5054e>.
- [33] G. Abd, A. M., and E.-S. M., “A Comparative Study of Meta-heuristic Algorithms for Solving Quadratic Assignment Problem,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 5, no. 1, pp. 1–6, 2014, doi: 10.14569/ijacsa.2014.050101.
- [34] J. Kennedy and R. Eberhart, “Particle Swarm Optimization,” *Stud. Comput. Intell.*, vol. 927, pp. 5–13, 1995, doi: 10.1007/978-3-030-61111-8\_2.
- [35] D. Bratton and J. Kennedy, “Defining a standard for particle swarm optimization,” *Proc. 2007 IEEE Swarm Intell. Symp. SIS 2007*, no. Sis, pp. 120–127, 2007, doi: 10.1109/SIS.2007.368035.
- [36] S. Rukhaiyar, M. N. Alam, and N. K. Samadhiya, “A PSO-ANN hybrid model for predicting factor of safety of slope,” *Int. J. Geotech. Eng.*, vol. 12, no. 6, pp. 556–566, 2018, doi: 10.1080/19386362.2017.1305652.
- [37] I. Kusmarna, L. K. Wardhani, and M. Safrizal, “Course Scheduling Application Using Particle Swarm Optimization Algorithm (PSO),” *J. Tek. Inform.*, vol. 8, no. 2, pp. 1–8, 2015, doi: 10.15408/jti.v8i2.2441.
- [38] S. Karkheiran, A. Kabiri-Samani, M. Zekri, and H. M. Azamathulla, “Scour at bridge piers in uniform and armored beds under steady and unsteady flow conditions using ANN-APSO and ANN-GA algorithms,” *ISH J. Hydraul. Eng.*, vol. 00, no. 00, pp. 1–9, 2019, doi: 10.1080/09715010.2019.1617796.
- [39] T. Qin, S. Zeng, and J. Guo, “Robust prognostics for state of health estimation of lithium-ion batteries based on an improved PSO-SVR model,” *Microelectron. Reliab.*, vol. 55, no. 9–10, pp. 1280–1284, 2015, doi: 10.1016/j.microrel.2015.06.133.
- [40] X. Yang, “Nature-Inspired Optimization Algorithms,” Elsevier, pp. 77–87, 2014, doi: 10.1007/SpringerReference\_72296.
- [41] L. Chato, S. Tayeb, and S. Latifi, “A genetic algorithm to optimize the adaptive Support Vector Regression model for forecasting the reliability of diesel engine systems,” 2017 IEEE 7th Annu. Comput. Commun. Work. Conf. CCWC 2017, vol. 1301726, pp. 0–6, 2017, doi: 10.1109/CCWC.2017.7868462.
- [42] F. Gorunescu, *Data Mining: Concepts, Models and Techniques*, INTELLIGEN. Springer, 2011.
- [43] B. K. Khotimah, M. Miswanto, and H. Suprajitno, “Optimization of feature selection using genetic algorithm in naïve Bayes classification for incomplete data,” *Int. J. Intell. Eng. Syst.*, vol. 13, no. 1, pp. 334–343, 2020, doi: 10.22266/ijies2020.0229.31.
- [44] D. Zhang, W. Liu, A. Wang, and Q. Deng, “Parameter Optimization for Support Vector Regression Based on Genetic Algorithm with Simplex Crossover Operator,” *J. Inf. Comput. Sci.*, 2011, [Online]. Available: [https://www.researchgate.net/publication/266051142\\_Parameter\\_Optimization\\_for\\_Support\\_Vector\\_Regression\\_Based\\_on\\_Genetic\\_Algorithm\\_with\\_Simplex\\_Crossover\\_Operator](https://www.researchgate.net/publication/266051142_Parameter_Optimization_for_Support_Vector_Regression_Based_on_Genetic_Algorithm_with_Simplex_Crossover_Operator).
- [45] J. Wu and Y. Xie, “Hybrid Support Vector Regression with Parallel Co-Evolution Algorithm Based on GA and PSO for Forecasting Monthly Rainfall,” pp. 524–539, 2019, doi: 10.4236/jsea.2019.1212032.

# Experimental Analysis and Monitoring of Photovoltaic Panel Parameters

Zaidan Didi, Ikram El Azami

Computer Science Research Laboratory (LaRI)-Faculty of Sciences,  
Ibn Tofail University, Kenitra, Morocco

**Abstract**—In this article, we establish a technique based on the internet of things to simultaneously monitor the main values that characterize a photovoltaic solar panel. This technique allows to discover the problems and the monstrosities during the operation. This study also allows to collect the parameters and quantities measured for analysis. This method is based on exploiting the advantages of IoT technology. For this it will be a good choice to use and exploit the Esp32 microcontroller, because the two WIFI and Bluetooth modules are integrated. The design process began by creating a system to measure the intensity of the electric current delivered by the photovoltaic panel. A current sensor was implemented for this purpose. To prevent damage to the microcontroller, a voltage divider was proposed to decrease the voltage at the pin level of the Esp32 for measurement. Next, the power and energy values were calculated to estimate the production capacity. In the final stage, a low-power Bluetooth link was created to transmit the four quantities to a smartphone or other compatible device. Real-time values were presented as graphs on the free ThingSpeak platform and displayed on both, an LCD screen and the serial monitor of the Esp32 microcontroller. The system was tested without any problems or errors.

**Keywords**—Current sensor; bluetooth low consumption; photovoltaic panel; Esp32 microcontroller

## I. INTRODUCTION

Currently, the exploitation of renewable energies is experiencing an intense and remarkable increase, to reduce and minimize costs and expenses in the energy field [1]. In this realization we strongly noted this last report between the electric energy and the energy generated by the photons coming from the sun, this report exceeds 25% [2], [3].

The proper functioning of solar panels is sensitive to being affected by climatic problems. As a result, the distortion negatively influences and can reduce the performance of photovoltaic panels [4], [5], [6], [7], [8]. To regularly detect the operating problems of photovoltaic systems, proactive management is necessary to ensure real-time monitoring of the values of the main parameters of this system. In this article, a design has been processed to measure and supervise the values of the quantities that characterize a photovoltaic panel, which will then be transmitted through low-consumption Bluetooth connectivity. Our document is ordered as follows. In the first section we have elaborated a summary. In Section II, similar studies have been presented which have dealt with the same objectives. Section III is reserved for the hardware part in which the essential characteristics of the Esp32 microcontroller and the ACS712-30A current sensor have been exposed. In the

same section, we have exposed a detailed conception of our realization. Section IV brings together the results and the discussion via diagrams and curves which represent the results obtained experimentally. Noted here that the values of the parameters are collected and stored on the ThingSpeak platform which is reserved for connected objects while we exploit the HTTP protocol. Finally, in the last section a conclusion closes this study.

## II. RELATED WORK

### A. Transfers Data with Bluetooth Modules, HC-06 and HC-05

In recent years and thanks to the technological revolution in the field of the Internet of Things, several scientific studies have integrated the Arduino microcontroller with the two external Bluetooth modules HC-05 and HC-06 [9], [10], [11], [12], [13]. These Bluetooth modules are greedy because they consume a large amount of energy, and we have noticed the non-compatibility with the Arduino card, which in many cases complicates the proper functioning of these modules to achieve connectivity via Bluetooth.

### B. Use of the F031-06 Voltage Sensor for the Measurements

In the same design, other delicate researches have used the F031-06 voltage sensor as the main element to measure the potential difference that characterizes the photovoltaic panel [14], [15], [16], [17], [18], thus, the operating principle of this sensor is essentially based on the structure of a resistive divider, therefore, we quickly concluded that this sensor is characterized by the non-stability of the measurements, moreover, over a given measurement range the output of this sensor is always a linear function, in the research we have done, we have noticed a wide sensitivity to noise and the values measured with this sensor are slightly different to the real values.

### C. Transfers Data with WIFI Module

In the same context, other research has integrated the WIFI module to transmit the different measured quantities of photovoltaic panels such as current, voltage and energy [19], [20], [21], [22], [23]. Noted here that in the studies used by the integrated WIFI module of the Esp32 microcontroller, the IP address is assigned dynamically by a DHCP server, so this address changes its value at each restart, this becomes a major problem and complicates the situation in the case where a web server is installed on the Esp32 microcontroller. In most cases, this inconvenience is overcome by integrating the Wifi.config function to set the IP address, this function accepts the

following attributes, Ip, gateway, dns, and subnet. In general, these studies are very successful, while the Wi-Fi network coverage space remains a challenge.

#### D. Arduino Mega as Main Processing Element

In other scientific research, the hardware design consists of an Arduino Mega board, therefore an ATmega2560 microprocessor [24], [25], [26], [27], [28]. These studies profusely reduce the hardware complexity and make the model clearer, We noted here that the Arduino Mega2560 board is characterized by several advantages such as the number of input / output (54 pins), a large number of analog pins (16 pins), serial ports (3 ports), an I2C port and an SPI port, compared to the ordinary Arduino board, the Mega2560 is characterized mainly by a large memory predisposition, let's also add compatibility with the majority of Arduino modules. the IoT.

### III. MATERIALS AND METHODS

In this section, the material design used in this realization is presented, it is based mainly on the following elements:

#### A. Microcontrôleur Esp32

In this study, the Esp32 Microcontroller plays a key role, because it is created to promote and ensure the learning and development of connected objects, and embedded applications [29], it is characterized by a large computing capacity in comparison to the Arduino, this microcontroller wakes up regularly when a specific condition is verified, In practice, experiments show that the Esp32 microcontroller greatly minimizes the amount of energy consumed thanks to the reduced duty cycle, The ESP32 microcontroller integrates the WROOM-32 microprocessor (Tensilica Xtensa LX6) which particularly works under a clock frequency of 240 MHz. Also noted that the Bluetooth (Bluetooth Low Energy) and WIFI modules (WIFI: 802.11 n width 2.4 GHz) are integrated.

#### B. ACS712 30a Current Sensor

The ACS712 current sensor is essentially based on the ascending Hall effect. Whatever the nature of the electric current (direct (DC) or alternating (AC)), the ACS712 sensor is connected in practical embodiments in series with the load, with a sensitivity of around 66mV per ampere, see Fig. 1, [30][31].

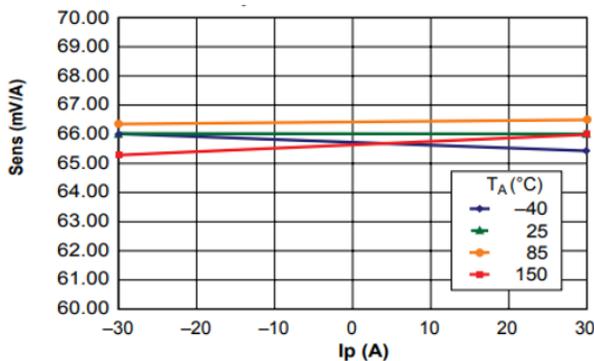


Fig. 1. Sensitivity of the ACS71230-a sensor

Among the advantages of the ACS712 sensor, the authors have noted that this sensor accepts the passage of electric current in both directions, this makes it possible to measure a positive or negative electric current in specific situations. The current inside this sensor generates a magnetic field; at the output, a DC voltage proportional to the current will be obtained. It is equally important to note that a calibration with a blank reading is strongly recommended. Note also that an external magnetic field (like for example) can negatively influence the measurements made. In practical implementations, the ACS712 Hall effect current sensor (10nF) is distinguished by intense noise corresponds to 130mA, the authors have greatly limited this noise by adding a 470nF capacitor, also have profusely limited this noise by adding a 470nF capacitor. The pins of the ACS712 are denoted in Fig. 2, the technical characteristics are given as follows: Dimensions: 31x13x15mm, Chip: ACS712ELEC-30A, Measured current range: -30A to +30A, Vref at 0A: Vcc/2 i.e. 2.5V, Sensitivity: 66mV/A, Insulation: 2.1KV, Consumption: 10mA, Error: 1.5 % at 25°C, Power supply: 5VDC (4.5-5.5VDC).

#### C. Autres Elements

Since the photovoltaic panels in our realization generate a maximum voltage equal to 42V, and since the voltage at the terminals of the ESP32 microcontroller must not go above 5v, next, we installed a voltage divider to lower the voltage to avoid unbuttoning the ESP32 microcontroller. The value of resistors R1 and R2 are calculated to identify the two conditions (the maximum voltage at the microcontroller terminals must not exceed 5V and the maximum voltage of the PVs equal to 42V) ie R1=1kΩ, R2=12KΩ. The voltage value at pin R2 is measured by ESP32 pin number 34 see Fig. 4. Other hardware elements are used, an LCD display to display the measured quantities like the current intensity value, voltage, power, energy. Finally, we have a rectangular photovoltaic panel with an area equal to 1.6 m<sup>2</sup>.

#### D. Comparison of Wireless Techniques

Table I, summarizes a comparison between five standard wireless technology, Bluetooth, RF 433MHz, ZigBee, and WIFI, therefore, the association standards, power consumption, bandwidth and throughput are represented in table1[32].

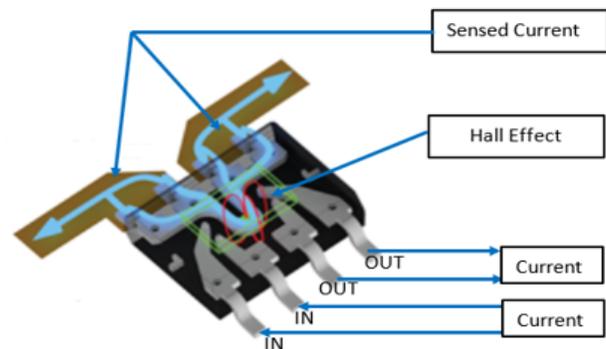


Fig. 2. Designation of the pins of the ACS712 sensor

TABLE I. COMPARISON BETWEEN STANDARD WIRELESS TECHNOLOGIES

Technologies	RF	WI-Fi	BlueTooth	ZigBee
Standards Association	IEEE 802.15.4	Low	433 MHz	4-9 Kb/s
Energy consumption	IEEE 802.11	High	2.4GHz	---
Frequency	IEEE 802.15.1	Low-High	2.4GHz	<24 Mbps
Debit	IEEE 802.15.4	High	2.4GHz	<1Mb/s

E. Formula Evaluation

To carry out this study, we chose an open area characterized by the absence of obstacles which can negatively influence the propagation of UHF waves, this place is also characterized by the absence of radio emissions on the 2.4 GHz frequency so no interference phenomenon, the weather conditions are very favorable with a very clear sky and a temperature of 29C. At first, we studied the aspect of propagation of UHF radio waves which constitutes a principle of bidirectional exchange for Bluetooth technology on a frequency band of 2.4 GHz, therefore, we studied the attenuation of these UHF waves to reduce obstacles and unwanted effects that can negatively influence the UHF signal coverage area. Note here that the undesirable epiphenomena of Ultra High Frequency wave attenuation are defined by Eq. (1) in [33], [34].

$$PL_{dB} = 10 \log_{10} \left( \frac{p_t}{p_r} \right) \quad (1)$$

- PLdB : Wave attenuation in dB,
- pt : Transmitted power
- pr : received power

Note that the distances related to attenuation are well determined, Fig. 3 presents this attenuation corresponding to ultra-high frequency waves (UHF)

In practice, in the UHF signal damping equation in dB, a coefficient is added to determine a true ratio between the acquired power P and the separation distance d, this coefficient is denoted by n and equation (2) in [36] represents this mutation (n: path factor , n=2 free space, X<sub>δ</sub>: Gaussian random variable, δ: standard deviation).

$$PL(d) = x_{\delta} + PL(d_0) + 10n \log_{10} \left( \frac{d}{d_0} \right) \quad (2)$$

In the field of telecommunications and signal transmission, the Friis Eq. (3) explains an interesting relationship between the distance d between the transmitter and the receiver in a free space, as well as the power and the antenna gain [37].

$$\frac{p_r}{p_t} = G_t G_r \frac{\lambda^2}{(4\pi d)^2} \quad (3)$$

- P<sub>t</sub>: Power delivered by the transmitter (W)
- P<sub>r</sub>: Power delivered by the receiver (W)
- G<sub>r</sub>: Receiver Gai

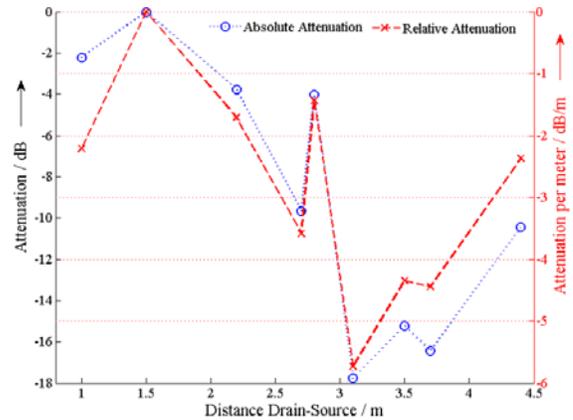


Fig. 3. Attenuation of UHF wave determined in time domain [35]

- G<sub>t</sub>: Transmitter Gain
- λ: Signal Wavelength
- d: Distance between (transmitter, receiver).

Scientifically, an important interpretation allows to exchange enormously to believe that the free space attenuation is proportional to the square of the frequency. Therefore, these expressions appear only in formulas where the antenna gain expression is represented. On the other hand, this help has totally disappeared the condition where we have antennas with a fixed surface. [38]. It is therefore concluded that the antennas are the cause of a loss of gain due to non-compliance. Equation (4) represents this effect.

$$\frac{p_r}{p_t} = G_t G_r (1 - |s_{11}|^2)(1 - |s_{22}|^2) \left( \frac{\lambda^2}{(4\pi d)^2} \right) \quad (4)$$

- s<sub>11</sub>/s<sub>22</sub>: reflection coefficients on the transmit/receive antenna.
- Gr/Gt: Receiver/Transmitter Gain
- λ: Signal Wavelength
- d: Distance between (transmitter, receiver).

F. The Main Design Proposal

In this section, we will develop the main diagram of our realization, as well as the advantages which draw the strong point of this monitoring system.

- The design and implementation are very easy.
- The use of data transfer via a low-consumption Bluetooth link and very useful in places far from communication networks (absence of GSM and Internet for example).
- Wireless monitoring – secure and reassured data exchange.

We have proposed the following general scheme for measuring current and voltage, in order to calculate power and energy, see Fig. 4. These four values will be sent via a low-power Bluetooth link that we must create in the next section.

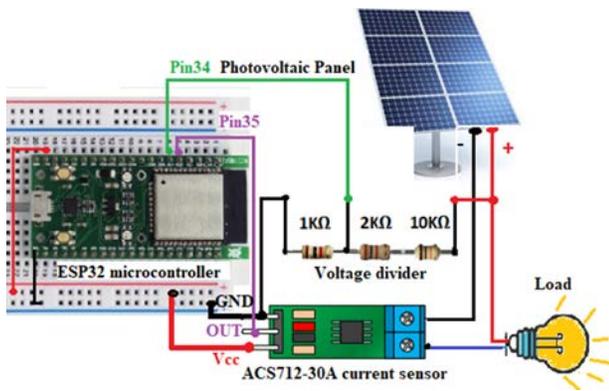


Fig. 4. Main design diagram

In this section, the main diagram of this study is proposed. It was started by measuring the current intensity with the ACS712-30A sensor at the level of pin35 of the Esp32 microcontroller, see Fig. 4. After an aid to lower the potential difference was developed, thanks to a voltage divider so as not to exceed the 5v threshold which represents the maximum voltage that the Esp32 microcontroller can support. The measurement is therefore carried out at the resistance level R2, therefore, the pin34. Note that to know the voltage between the terminals of the photovoltaic panel, the mesh law and Ohm's law was used. And finally, the values of the voltage (V) and the current (I) we used to calculate in an exact way the two other parameters: the instantaneous power (P) and the energy (E).

#### G. Creation of the Bluetooth Link

The classic HC-05 and HC-06 external modules are outdated despite the simplicity of integration in IoT projects, these two modules have major drawbacks, such as large power consumption, another thing is that these two modules use the old Bluetooth V2.0, this is a heavily outdated version. To remedy this problem linked to high energy consumption, the Esp32 was used because the Bluetooth module is integrated into the body of this microcontroller by the manufacturer. The choice to use an Esp32 is highly difficult because it is characterized by a flow rate of around 1Mb/s and the consumption will be reduced 10 times thanks to the new cell construction technology [39]. In the main code, to establish and ensure a connection via Bluetooth Low Energy, first the BluetoothSerial.h library was declared; this library comprises all the functionality necessary to guarantee the serial connection via Bluetooth technology. Subsequently, the declaration of a BluetoothSerial class object is mandatory, this object is called SerialBT.

To initialize the Bluetooth module, it is must to interact with this SerialBT object. To listen to the client connection event, in the second step the callback function and the void setup() function should be configured. Still in the code, started the serial communication with a rate equal to 115200 baud, naming the begin() method on the BluetoothSerial object allows us to initialize the Bluetooth interface. It receives the name of the Esp32 microcontroller as input, in our realization "ESP32\_iosm". Finally in the void loop() function, a verification of the reception of data via the Bluetooth module. The working process of our study is shown in Fig. 5.

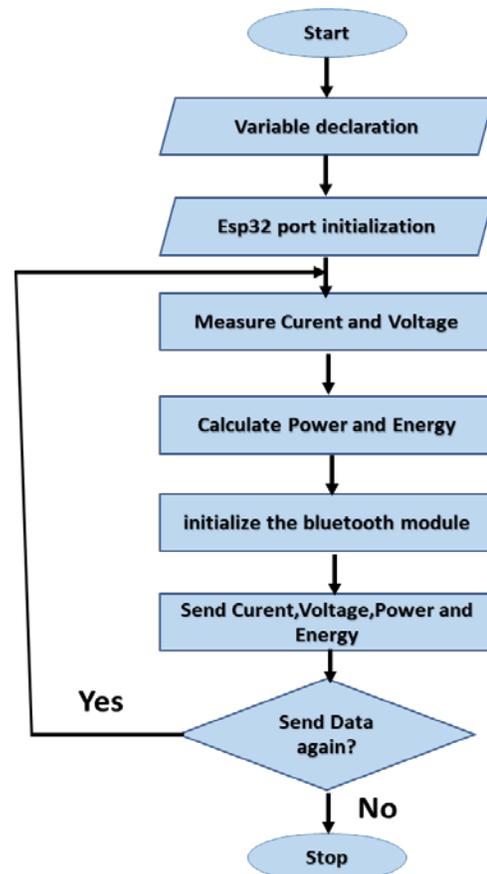


Fig. 5. Operation process flowchart

#### IV. RESULTS AND DISCUSSION

This achievement has been implemented correctly, see Fig. 6 , the transfer of data via this Bluetooth link is a good affirmation of this success, the screenshots in Fig. 8 and Fig. 9 express and describe in real time the quantities transmitted ( Energy, Power, Voltage and Current).

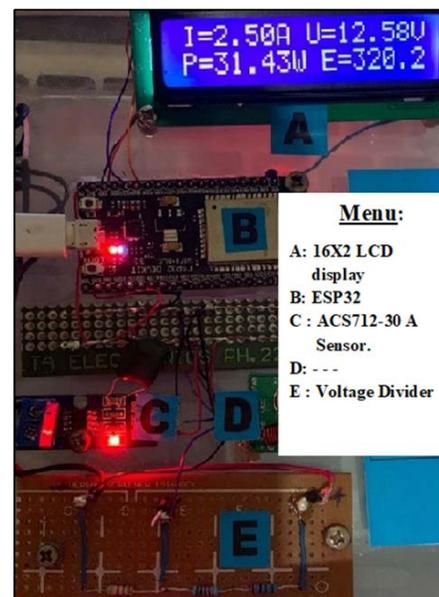


Fig. 6. Practical realization of the study[29]

A. Experimental Results

The Arduino Integrated Development Environment (IDE) was used to compile and upload the code, our Esp32 microcontroller will be immediately recognized and detected by other Bluetooth devices under the name we assigned in the previous section “ESP32\_iosm”, see Fig. 7.

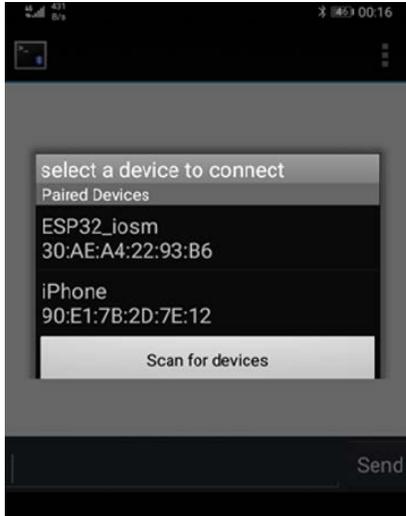


Fig. 7. Detection of the Esp32 microcontroller on smartphone

A new COM port will be available to perform the pairing when the pairing action is completed. After a discovery operation, Fig. 8 exposes on a smartphone the data received as a result of a low-power Bluetooth communication.



Fig. 8. Receiving data on a smartphone

The different experimental results are grouped in Table II, these results imprint the first five loops of this achievement, these values are displayed in real time on the serial monitor of our microcontroller, see Fig. 9.

TABLE II. MEASUREMENT- EXTRACT

Loop number	Current (A)	Voltage (V)	Power (W)	Energy (Ws)
1	1.53	14.59	22.33	102.70
2	2.68	14.63	39.29	141.99
3	1.53	14.62	22.49	164.42
4	1.46	14.60	21.37	185.79
5	1.52	14.60	22.17	207.96

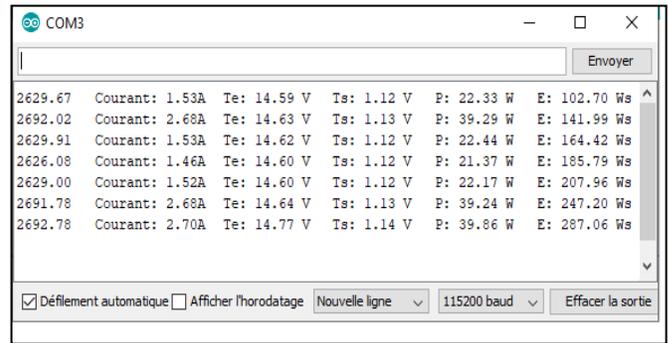


Fig. 9. Preview of the result on the serial monitor

Finally, to facilitate the comparison and make the data clearer in the form of representative graphs, we made use of a free "open source" web application, the TingSpeak platform, which is reserved for the realization of the Internet of Things and the embedded electronics projects, Fig. 10, we used the http protocol to collect the data to transfer the connected objects. In the Fig. 11, we have presented the variations of energy / second and Fig. 12 displays a representation of the total energy of our photovoltaic panel.

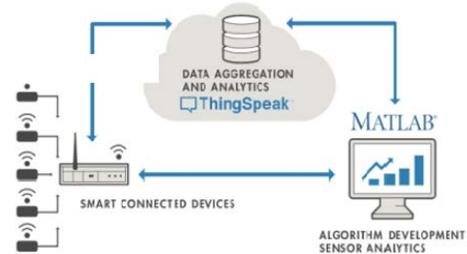


Fig. 10. TingSpeak plateforme

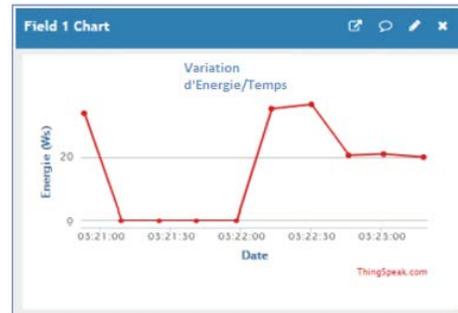


Fig. 11. Energy variations/second



Fig. 12. Energy total variation

## V. CONCLUSION

The experimental results of this study show that monitoring the main parameters of photovoltaic panels with low-power Bluetooth modules is still a good choice in isolated places, or when GSM or Internet network coverage is absent. Note that the implementation of these projects is simple, easy and not expensive. The monitoring of these parameters in real time and in distinct intervals allows us to discover and see operating strangeness. This greatly increases the credibility of the system and guarantees its proper functioning.

## REFERENCES

- [1] Ahmed A, Al-Amin AQ, Ambrose AF, Saidur R (2016) Hydrogen fuel and transport system: a sustainable and environmental future. *Int J Hydrog Energy* 41:1369–1380.
- [2] K. Shanmugam, G. Pitto and L. Barbato, "Gridcode compliances and Operational Requirements of Grid connected BESS-Renewable Power Plants," 2020 IEEE International Conference on Power Systems Technology (POWERCON), pp. 1-6 (2020).
- [3] NREL: Best Research-Cell Efficiency chart, <https://www.nrel.gov/pv/assets/pdfs/best-research-cell-efficiencies.20190802.pdf> (accessed: September 2019).
- [4] Alswidi, Mohamed, Abdulaziz Aldobhani, and Abdurraqib Assad. "Design and simulation of adaptive controller for single phase grid connected photovoltaic inverter under distorted grid conditions." *International Journal of Advanced Computer Science and Applications* 8.11 (2017).
- [5] Jamal, A., Putri, S. G., Chamim, A. N. N., & Syahputra, R. (2019). Power Quality Evaluation for Electrical Installation of Hospital Building. *International Journal of Advanced Computer Science and Applications*, 10(12).
- [6] Emar, W., Al-omari, Z., & Saraereh, O. A. (2019). Optimization of Cúk Voltage Regulator Parameters for Better Performance and Better Efficiency. *International Journal of Advanced Computer Science and Applications*, 10(11).
- [7] Kotsopoulos, A., Heskes, P. J., & Jansen, M. J. (2005). Zero-crossing distortion in grid-connected PV inverters. *IEEE Transactions on Industrial Electronics*, 52(2), 558-565.
- [8] Ayub, M., Gan, C. K., & Kadir, A. F. A. (2014, May). The impact of grid-connected PV systems on Harmonic Distortion. In 2014 IEEE Innovative Smart Grid Technologies-Asia (ISGT ASIA) (pp. 669-674). IEEE.
- [9] Gajendrasinh N. MoriPriya R. Swaminarayan, Measuring IoT Security Issues and Control Home Lighting System by Android Application Using Arduino Uno and HC-05 Bluetooth Module, *Data Science and Intelligent Applications. Lecture Notes on Data Engineering and Communications Technologies*, pp 375-382 vol 52. (2021), Springer, Singapore.
- [10] Mehta, S., Saraff, N., Sanjay, S. S., & Pandey, S. (2018). Automated agricultural monitoring and controlling system using hc-05 bt module. *International Research Journal Of Engineering And Technology (IRJET)*, 5(5).
- [11] Mori, G. N., & Swaminarayan, P. R. (2021). Measuring IoT security issues and control home lighting system by android application using Arduino Uno and HC-05 bluetooth module. In *data science and intelligent applications* (pp. 375-382). Springer, Singapore.
- [12] Ranjitha, B., Nikhitha, M. N., Aruna, K., & Murthy, B. V. (2019, June). Solar powered autonomous multipurpose agricultural robot using bluetooth/android app. In 2019 3rd International conference on Electronics, Communication and Aerospace Technology (ICECA) (pp. 872-877). IEEE.
- [13] Singh, T., & Thakur, R. (2019). Design and development of PV solar panel data logger. *International Journal of Computer Sciences and Engineering (IJCSE)*, 7.
- [14] El Hammoumi, A., Motahhir, S., Chalh, A., El Ghzizal, A., & Derouich, A. (2018). Low-cost virtual instrumentation of PV panel characteristics using Excel and Arduino in comparison with traditional instrumentation. *Renewables: wind, water, and solar*, 5(1), 1-16.
- [15] Anand, R., Pachauri, R. K., Gupta, A., & Chauhan, Y. K. (2016, July). Design and analysis of a low cost PV analyzer using Arduino UNO. In 2016 IEEE 1st international conference on power electronics, intelligent control and energy systems (ICPEICES) (pp. 1-4). IEEE.
- [16] Gupta, V., Raj, P., & Yadav, A. (2017, September). Design and cost minimization of PV analyzer based on arduino UNO. In 2017 IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI) (pp. 1337-1342). IEEE.
- [17] Khattab, O. T., Alshmmri, M. A., & Marie, M. J. (2022). Implementation and Design of a Monitoring System for Tikrit Substation Using IoT. *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, 13(03), 607-616.
- [18] Routray, D., Rout, P. K., & Sahu, B. K. (2021, October). Real-time implementation of the MPPT algorithm based on fuzzy logic for solar PV system. In 2021 International Conference in Advances in Power, Signal, and Information Technology (APSIT) (pp. 1-7). IEEE.
- [19] Sapaklom, T., Janhom, K., Sipirah, C., Kjitdamkean, P., Ayudhya, P. N. N., Mujjalinvimut, E., & Kunthong, J. (2022, November). IoT Based IV and PV Curve Analyzer system for small PV panels PART I. In 2022 25th International Conference on Electrical Machines and Systems (ICEMS) (pp. 1-4). IEEE.
- [20] Maguluri, L. P., Srinivasarao, T., Syamala, M., Ragupathy, R., & Nalini, N. J. (2018). Efficient smart emergency response system for fire hazards using IoT. *International Journal of Advanced Computer Science and Applications*, 9(1).
- [21] ZERARI H, MESSIKH L, KOUZOU A, OUCHTATI S, Implementation of Smart Energy Management and Monitoring System for Public Lighting System Based on Photovoltaic and Storage Systems, in *Electrotehnica, Electronica, Automatica (EEA)*, 2021, vol. 69, no. 1, pp. 20-28, ISSN 1582-5175.
- [22] Barik, L. (2019). IoT based temperature and humidity controlling using Arduino and raspberry Pi. *International Journal of Advanced Computer Science and Applications*, 10(9).
- [23] de Dios Fuentes-García, J., Flores-Arias, J. M., Bellido-Outeiriño, F. J., Quiles-Latorre, F. J., Ortiz-López, M. A., & Garrido-Zafra, J. (2019, September). Monitoring of photovoltaic systems for self-consumption without over-consumption. In 2019 IEEE 9th International Conference on Consumer Electronics (ICCE-Berlin) (pp. 239-241). IEEE.
- [24] Mahzan, N. N., Omar, A. M., Rimon, L., Noor, S. M., & Rosselan, M. Z. (2017). Design and development of an arduino based data logger for photovoltaic monitoring system. *Int. J. Simul. Syst. Sci. Technol*, 17(41), 15-1.
- [25] Gusa, R. F., Sunanda, W., Dinata, I., & Handayani, T. P. (2018, March). Monitoring system for solar panel using smartphone based on microcontroller. In 2018 2nd international conference on green energy and applications (ICGEA) (pp. 79-82). IEEE.
- [26] Priharti, W., Rosmawati, A. F. K., & Wibawa, I. P. D. (2019, November). IoT based photovoltaic monitoring system application. In *Journal of Physics: Conference Series* (Vol. 1367, No. 1, p. 012069). IOP Publishing.
- [27] Samara, S., & Natsheh, E. (2019). Intelligent real-time photovoltaic panel monitoring system using artificial neural networks. *IEEE Access*, 7, 50287-50299.
- [28] Sugiarta, N., Sugina, I. M., Putra, I. D. G. A. T., Indraswara, M. A., & Suryani, L. I. D. (2018, December). Development of an arduino-based data acquisition device for monitoring solar PV system parameters. In *International Conference on Science and Technology (ICST 2018)* (pp. 995-999). Atlantis Press.
- [29] Didi, Z., & El Azami, I. (2021, January). IoT design and realization of a supervision device for photovoltaic panels using an approach based on radiofrequency technology. In *International Conference on Digital Technologies and Applications* (pp. 365-375). Springer, Cham.
- [30] Allegro Microsystem. ACS712: fully integrated, hall effect-based linear current sensor with 2.1 kVRMS voltage isolation and a low-resistance current conductor (2017), ACS712-DS, Rev. 7. Northeast Cutoff Worcester, Massachusetts 01615-0036 U.S.A.

- <https://www.sparkfun.com/datasheets/BreakoutBoards/0712.pdf>,  
Accessed 25 Aug 2017.
- [31] Shafique, M. T., Kamran, H., Arshad, H., & Khattak, H. A. (2018, November). Home energy monitoring system using wireless sensor network. In 2018 14th International Conference on Emerging Technologies (ICET) (pp. 1-6). IEEE.
- [32] Prasad, D., Chiplunkar, N. N., & Nayak, K. P. (2017). A trusted ubiquitous healthcare monitoring system for hospital environment. *International Journal of Mobile Computing and Multimedia Communications (Ijmcmm)*, 8(2), 14-26.
- [33] Andersen, J. B., Rappaport, T. S., & Yoshida, S. (1995). Propagation measurements and models for wireless communications channels. *IEEE Communications magazine*, 33(1), 42-49.
- [34] Tariq, S. A. M. (2016). Characterization and modelling of scattered wireless channel at 60 GHz in an underground mine gallery. *Ecole Polytechnique, Montreal (Canada)*.
- [35] Coenen, S., Tenbohlen, S., Markalous, S. M., & Strehl, T. (2008, April). Attenuation of UHF signals regarding the sensitivity verification for UHF PD measurements on power transformers. In 2008 International Conference on Condition Monitoring and Diagnosis (pp. 1036-1039). IEEE.
- [36] Cui, P. F., Yu, Y., Lu, W. J., Liu, Y., & Zhu, H. B. (2017). Measurement and modeling of wireless off-body propagation characteristics under hospital environment at 6–8.5 GHz. *IEEE Access*, 5, 10915-10923.
- [37] Griffin, J. D., & Durgin, G. D. (2009). Complete link budgets for backscatter-radio and RFID systems. *IEEE Antennas and Propagation Magazine*, 51(2), 11-25.
- [38] Malviya, L., Panigrahi, R. K., & Kartikeyan, M. V. (2018). Four element planar MIMO antenna design for long-term evolution operation. *IETE Journal of Research*, 64(3), 367-373.
- [39] Abdelatty, O., Chen, X., Alghaihab, A., & Wentzloff, D. (2021). Bluetooth Communication Leveraging Ultra-Low Power Radio Design. *Journal of Sensor and Actuator Networks*, 10(2), 31.

# Hybrid Feature Selection Algorithm and Ensemble Stacking for Heart Disease Prediction

Nureen Afiqah Mohd Zaini<sup>1</sup>, Mohd Khalid Awang<sup>2</sup>

Faculty of Informatics and Computing, Universiti Sultan Zainal Abidin,  
22000 Tembil, Terengganu, Malaysia

**Abstract**—In cardiology, as in other medical specialties, early and accurate diagnosis of heart disease is crucial as it has been the leading cause of death over the past few decades. Early prediction of heart disease is now more crucial than ever. However, the state-of-the-art heart disease prediction strategy put more emphasis on classifier selection in enhancing the accuracy and performance of heart disease prediction, and seldom considers feature reduction techniques. Furthermore, there are several factors that lead to heart disease, and it is critical to identify the most significant characteristics in order to achieve the best prediction accuracy and increase prediction performance. Feature reduction reduces the dimensionality of the information, which may allow learning algorithms to work quicker and more efficiently, producing predictive models with the best rate of accuracy. In this study, we explored and suggested a hybrid of two distinct feature reduction techniques, chi-squared and analysis of variance (ANOVA). In addition, using the ensemble stacking method, classification is performed on selected features to classify the data. Using the optimal features based on hybrid features combination, the performance of a stacking ensemble based on logistic regression yields the best result with 93.44%. This can be summarized as the feature selection method can take into account as an effective method for the prediction of heart disease.

**Keywords**—Heart disease prediction; feature selection; stacking; accuracy

## I. INTRODUCTION

The process of learning a function that maps an input to an output based on examples of input-output pairs is referred to as supervised learning in the field of machine learning. This task involves learning a function that maps an input to an output. It accomplishes this by drawing conclusions about a function based on a collection of samples from training that have been labelled [1]. In a variety of fields, including marketing, commercial applications, pattern recognition, image processing, classification, and prediction, feature selection has been utilized. It is common to encounter a sizable data collection and a high number of features while working with actual applications. Most of the time, just a few of the features are important and pertinent to the objective. Since the remaining features are viewed as unimportant and unnecessary, doing without them would not only affect performance but also classification accuracy. As a result, choosing a suitable and compact feature subset from the original features is crucial to improving classification performance and accuracy as well as overcoming the curse of dimensionality. To determine the importance of attributes, feature selection techniques are employed, and the aim is to

minimize the number of input variables to those demands most relevant to the model. Aside from minimizing the number of attributes, feature selection also reduces processing time as well.

According to [2], medical records from the National Heart Institute Malaysia (IJN) discovered between January 1, 2009, and December 31, 2018, were used in a non-interventional study that looked back 10 years. From the IJN database, there were 3923 out of 4739 eligible and used in the analysis. Another study by [3] in 2019, conducted by the Department of Statistics of Malaysia, found that heart disease was the leading cause of death in Malaysia. Representing 15% of all fatalities requiring rapid medical attention. However, heart disease can be prevented by avoiding dangerous factors. In machine learning, varieties of algorithms such as supervised, unsupervised, semi-supervised, reinforcement, and transduction, are frequently employed. Supervised learning is the ability of an algorithm to synthesize knowledge from previously labelled data in order to predict future unlabelled cases [4].

In this study, 13 attributes from the UCI dataset are used for the experiment to determine the cause of heart disease. Nevertheless, not all attributes are useful, and a feature selection method is needed to prove the only important cause of heart disease. The choice of attributes based on the feature selection method might vary depending on the feature selection method used. The prediction of heart disease can be detected based on symptom from patients which make the specialist's task easier. When we talk about predicting heart disease, we should note that prediction is one of the applications of machine learning that is utilized frequently. With the assistance of machine learning, data mining is quickly becoming an essential part of the healthcare industry by employing classification and prediction techniques which are used to generate models that describe necessary classes [5].

It is commonly held risk factors such as age, sex, chest pain type, trestbps, chol, fasting blood sugar, restecg, thalach, exang, oldpeak, slope, number of major vessels, and thalassemia are the major risk factors for heart disease according to the dataset used. In light of these considerations, this research employed a feature selection method to build a heart disease risk assessment model that could aid specialists in making accurate early predictions [6].

Even though several feature selection strategies have been used in decision support systems for medical datasets, there is

always the opportunity for improvement. The combination of feature selection algorithms and classifiers has to be tuned for heart disease datasets with a lot of feature space in order to deliver high performance. The proposed framework is based on a well-balanced mix of two different types of feature selection algorithms that work well together.

This study aims to propose a hybrid feature selection that combines both chi-squared and ANOVA techniques. Chi-squared is utilized for the selection of categorical features, whereas ANOVA is applied to numerical features. The research proposes to combine the highest rank from both techniques, and the five most influential features are derived from a total of 13 features. The five most influential features are then evaluated using an ensemble stacking approach to improve the accuracy of heart disease prediction.

This paper is organized as follows: Section II discovered related works consisting of accuracy achieved by the author using feature selection technique for the prediction of heart disease. Section III discusses the dataset to use for the experiment along with the feature selection technique and framework that visualize the whole process for the experiment. Lastly, section IV discussed the result obtained based on the experiment made, and in Section V, the conclusion is presented.

## II. LITERATURE REVIEW

There are several elements that lead to heart disease, however the present approaches for heart disease prediction are inadequate and need to be improved. By using reduction approaches to remove some of the redundant features, the prediction accuracy might be improved. Feature selection is a process of selecting important attributes of the dataset. Pre-processing is the main step for selecting important attributes for a certain dataset. In this research, ten base classification algorithms and three subsets of meta-models are tested for the prediction of accuracy.

### A. Filter Method

The filter method is one of the feature selection methods that independently evaluates the importance of each feature. The selected features are subsequently used as input for a model-building process.

Before induction can take place, the filter method is used to remove unwanted attributes using one paradigm which independently act[7]. Karl Pearson pioneered the use of chi-squared statistics for categorical data, but it will take some time before the asymptotic distribution of these statistics was thoroughly understood [8].

However, the valid conclusion from chi-squared depends on several assumptions such as [9]:

- 1) A cross-tabulation can be used to figure out actual frequencies. The chi-squared test should not be used for percentages or other derived statistics.
- 2) The two variables are nominal which is the categories have no natural ordering.
- 3) Independent observations.

- 4) More than 75%-80% of contingency table columns have an expected count of  $\geq 5$ , and none have an expected count of 0.

Aside from chi-squared, [10] ANOVA test is another filter-based feature selection technique used in this research. By utilizing the SelectKBest class, the `f_classif()` function is called upon to determine the most important features. SelectKBest class may be found in the scikit-learn library which employs a scoring function to assign the features with the highest score.

According to [11], Classification and Regression Trees (CART), Gradient Boosting Machine (GBM), Adaboost, K-Nearest Neighbor (KNN), Multilayer Perceptron (MLP), Stochastic Gradient Descent (SGD), Support Vector Classifier (SVC) and Naïve Bayes are tested through feature selection to find the best accuracy algorithm. CART was found to have the best accuracy with 87.65%. Four important attributes from eleven features are selected based on the feature selection. The author uses the majority voting technique to find out the best attributes and the result proved that `st_slope_flat` and `st_depression` are the best and second highest results go to `max_heart_rate_achieved`, `exercise_induced_angina`, and `cholesterol`. The authors claimed these attributes are the leading cause of heart disease.

In 2019 [12], the author uses a rapid miner as a tool to test the accuracy of each algorithm. Six algorithms such as Decision Tree, Logistic Regression, Logistic Regression SVM, Naïve Bayes, and Random Forest are tested and the result found out Logistic Regression SVM is the highest with 84.85%. However, the author did not reveal the attributes of the leading cause.

On the other hand, [13] investigated the use of principal component analysis (PCA) in clinical aspect. To measure the effectiveness of reducing infection risk among university student, a pilot study of 200 volunteers was carried out. Essential clinical parameters were identified and confirmed by medical experts. From the clinical history variables with 49 parameters, the disease was identified through the use of PCA. PCA method was utilized to confirm the weightage of risk level towards the disease in order to ensure the system possesses the highest possible level of accuracy, reliability, and efficacy. Cumulative achieved with the use of PCA is 58.288% and the author proof optimal accuracy, reliability, and efficiency to conduct mass-screening of students.

The author in [14], found the most accurate algorithm achieved 85.00% using chi-squared feature selection with the BayesNet classifier. The dataset of heart disease is tested using principal component analysis (PCA), chi-squared testing, ReliefF, and symmetrical uncertainty. The author agreed to use PCA feature extraction with IBK and the result is highest for recall at 87.22% but the accuracy is low compared to the chi-squared result. Based on the results, `cp` is categorized as the most influential feature for heart disease prediction followed by `exang`, `chol`, and `thal`. Different features are ranked differently based on which feature selection is used.

Based on the dataset, this research [15] compares several machine learning techniques and determines the most efficient classification technique. KNN, NB, decision tree (J48), and RF are four different classification algorithms and other techniques, such as SVM were used to compare with affinity degree (AD) classification. All these algorithms are then tested on three different UCI dataset. As a result, J48 demonstrates the highest level of performance when compared to the other four classifiers as the purpose of this research is to investigate the compatibility of affinity to use for classification method.

The study by [16] affirms the use of the backward feature selection technique resulted in the highest accuracy of 88.52% using the decision tree algorithm. Algorithms such as random forest, support vector machine, decision tree, k-nearest neighbor, logistic regression, and gaussian naïve bayes are tested and the decision tree outperformed the other five algorithms. They also experimented with the accuracy using ten different feature selection techniques which are ANOVA, chi-squared, mutual information, ReliefF, forward feature selection, backward feature selection, exhaustive feature selection, recursive feature elimination, lasso regression, and ridge regression. As a result, backward feature selection is the most influential feature selection technique which leads to a better result.

Research done by [17], suggested dataset of 70000 patients and 11 features are tested with the chi-squared feature selection method. Features involved in this research consist of age, gender, height, weight, systolic blood pressure, diastolic blood pressure, cholesterol, glucose, smoking, alcohol intake, and physical activity. Seven algorithms and the chi-squared method were used to filter the most influential features. The author adjusted some features of the dataset to discover the factors that have the greatest impact on cardiovascular disease which resulted in weight and height as the most influential cardiovascular disease. As a result, Multi-Layer Perceptron achieved the highest accuracy with 87.23%.

The authors of the research [18], proposed two different datasets and use a feature selection technique to find the best features. The author also tested the ensemble classifier with a sampling technique to find the best accuracy. ANOVA is one of the feature selection techniques used by the author to find the best features for improved accuracy. The study [19] suggested a model predict numerous diseases as there are very few suggestions made about the detection of numerous diseases. The author takes into consideration conditions such as heart disease, diabetes, and kidney disease. There are only a few features in the dataset that will not affect how well the prediction system works and only important features will be taken into consideration for the decision-making. Chi-squared and ANOVA are applied to trace out the best features from the dataset. Exang, cp, ca, oldpeak and thalach are chosen as the most influential features.

A study conducted by [20], shows the size of the dataset increase as the complexity of the model increases. Classification and regression fields are tested respectively in this research for comparison purposes as they might be a potential resource for the researcher to decide on appropriate

algorithms. Chi-squared as one of the feature selection methods is used for categorical, ordered with missing values, and ordered without missing values. The major benefit of chi-squared is, it decreases computing complexity through the merging procedure by decreasing the number of categories for each predictor.

Recently, [21] developed a heart failure survival prediction model with the help of an ensemble tree machine learning approach. Extreme Gradient Boosting (XGBoost) was demonstrated as the most accurate classifier with 83.00%. During the pre-processing stage, the unimportant feature will be removed to obtain better accuracy. The author uses ANOVA and chi-squared to analyze numerical and binary features, respectively. The most influential features consist of anemia, time, ejection\_fraction, and serum\_creatinine but 'time' features are counted as the highest contribution for the improvement of accuracy.

A comparison of the result obtained shows that different authors came out with different results. The highest accuracy achieved based on past work is 88.52% from the decision tree. Thal features can be categorized as the most influential features seems all the experiments conducted with feature selection show that thal ranked the most among other features. As will be shown in succeeding sections, we analyze and present a comparison with our feature selection technique together with the result of accuracy for heart disease prediction.

### III. METHODOLOGY

This study is based on the UCI dataset of heart disease which consists of 303 datasets and 13 attributes. The original attributes consist of age, sex, cp, trestbps, chol, fbs, restecg, thalach, exang, oldpeak, slope, ca, thal, and target. Data dictionary in Table I will explain further the attributes involved.

TABLE I. DATA DICTIONARY FOR INVOLVED FEATURES

Feature Name	Data Description
X <sub>1</sub> = Age	age
X <sub>2</sub> = Sex	1=male, 0=Female
X <sub>3</sub> = Cp	chest pain type: Value 0=typical angina, Value 1=atypical angina, Value 2=non-anginal pain, Value 3=asymptomatic
X <sub>4</sub> = Trestbps	resting blood pressure
X <sub>5</sub> = Chol	serum cholesterol in mg/dl
X <sub>6</sub> = Fbs	(fasting blood sugar > 120) 1=True, 0=False
X <sub>7</sub> = Restecg	resting electrocardiographic results: Value 0=normal, Value 1=having ST-T wave abnormality, Value 2=showing probable or definite left ventricular
X <sub>8</sub> = thalach	max heart rate achieved
X <sub>9</sub> = Exang	exercise induced angine: 1=yes, 0=no
X <sub>10</sub> =Oldpeak	ST depression
X <sub>11</sub> = Slope	slope of peak exercise:Value 0=upsloping, Value 1=flat, Value 2= downsloping

$X_{12}$ = Ca	number of major vessel(flourosopy)
$X_{13}$ =Thal(Thalassemia)	0 = error (in the original dataset 0 maps to NaN's),1 = fixed defect,2 = normal,3 = reversable defect
Y= Target	0 = no disease,1 = disease

Cases of heart disease and non-heart disease are extracted from the dataset and displayed visually.

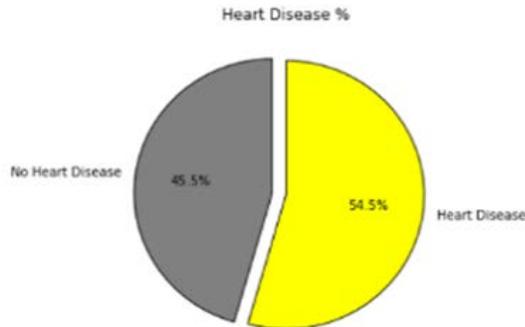


Fig. 1. Data visualization of heart disease and non-heart disease patients

According to Fig. 1, 54.50% of patients suffer from heart disease and the remaining 45.50% are free from heart disease.

The execution is accomplished using the following procedures:

- 1) UCI Cleveland dataset is obtained
- 2) Data visualization is performed
- 3) Dataset is divided into testing and training data
- 4) Applying algorithms method for training
- 5) Train the model
- 6) Heart disease prediction based on accuracy obtain

From the UCI dataset, 80% of the dataset has been assumed as training input for machine learning methods, and the model has been fitted accordingly. The remaining 20% is test data for predicting heart disease [22].

### B. Pre-Processing

Dimensionality reduction is a pre-processing procedure that can eliminate irrelevant data, noise, and redundant features to improve the accuracy of learning features and save training time [23]. Data pre-processing often encompasses the following task [24]:

**Data cleansing:** The first stage in data cleansing is identifying mistakes and inconsistencies in the database by evaluating the data. In other words, this phase is known as data audits and will identify all forms of database irregularities [25].

**Normalization:** Initially, pre-processing is not only a method for transforming raw data into a clean dataset but it also improves the performance of machine learning. By way of explanation, if data is acquired from various sources, it is collected in a raw format which is incompatible with analysis and machine learning [26].

**Feature discovery:** Feature discovery is one of the pre-processing methods which is the data filtered from the pre-

processing section. The advantage of feature discovery is extracting meaningful data from identified correlations of patterns [27].

**Management of imbalance data:** An example of an issue known as imbalance data classification is when the proportional class size of a dataset differs significantly by a significant margin from one another. From this, a group of a small number is represented as a minority class and the remaining belong to the other group represented majority class [28].

### C. Feature Selection

Attribute or feature selection is a data reduction method that is applied to the dataset. This method decreases the size of the data by eliminating unnecessary or duplicate attributes. Methods for selecting features subset can be broken into four distinct categories which are the embedded method, wrapper method, filter method, and hybrid method [29]. In our research, the features are divided into numerical and categorical which is the filter method applied. As it operates independently from the induction algorithm, this method is faster than the wrapper approach and produces a better generalization. However, the chi-squared method favours selecting a subset with a large number of features, necessitating a threshold to select a subset [30].

According to [31], it is found that the filter approaches are effective, scalable, computationally straightforward, and independent of the classifier. In this research, categorical features consist of sex, fb, restecg, exang, slope, ca, thal and target while age, trestbps, chol, thalach, and oldpeak are numerical features. Chi-squared is used to generate categorical features and ANOVA is tested for numerical features. Both methods generate the features according to rank based on the importance of each feature. The Table II below shows the selected features for categorical and numerical features.

TABLE II. SELECTED FEATURES FOR CATEGORICAL AND NUMERICAL

Feature Selection	Selected Features
Categorical	$X_{12}, X_3, X_9$
Numerical	$X_8, X_{10}$

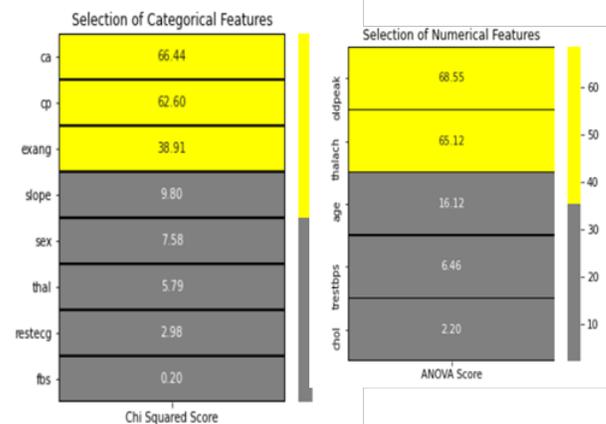


Fig. 2. Ranked each feature selection method for categorical and numerical features

Fig. 2 depicts the ranking of the involved features based on their importance. Ca, cp, and exang counted as the highest rank tested with chi-squared for categorical features while oldpeak and thalach counted as the most influential features for the numerical group using the ANOVA score.

#### D. Chi-Squared

Chi-squared is one of the techniques for categorical types of data. The chi-squared test determines if two categorical variables are significantly associated. Two-sided chi feature selection is tested between each categorical and binary outcome with a p-value. The features are retained with two-sided  $p < 0.05$  [32].

Several steps involved in the chi-square process are explained as follows [33]:

Step 1: All features from the original dataset are selected.

Step 2: Utilize the chi-squared () function from the scikit-learn to figure out whether the two features are independent or not. Use (1) to find the chi-squared score for each of the following features.

$$X^2 = \sum \frac{(f_o - f_E)^2}{f_E} \quad (1)$$

Step 3: The value with the highest chi-squared value probably relies on the target feature and is therefore selected for model creation. SelectKBest() was utilized to choose the five features with the highest chi-squared value.

Step 4: The next step is to determine a threshold to construct a subset for the number of features represented by n. The optimal number of features with the highest Chi test score is utilized based on the top five ranking features. In this research, five features with the highest Chi test score are tested to create the original feature subset.

According to [34], the strategy for the chi-squared method is incrementally adding important characteristics to the feature subset. At each level, this method will determine the significance threshold and discards features that fall below it. As a result, the chi-squared strategy is more efficient than similar step-wise selection methods. Most of the studies prove the use of the chi-squared method among other feature selection methods improves most of the classifiers' performance and accomplishes outstanding results [35].

Based on [36], up to 1900, the evolution of the chi-squared test process can be divided into six stages. Six related stages included:

- 1) From the multivariate error law to the multivariate normal distribution.
- 2) Exponent distribution in multivariate normal density.
- 3) Multinomial distribution approximation by multivariate normal density.
- 4) Evaluation of the exponent when the moment is multinomial.
- 5) The definition to which probability refers.
- 6) Provision for the effect of estimating an undetermined parameter.

#### E. ANOVA

Analysis of Variance (ANOVA) is another technique used for the classification method. ANOVA is tested for numerical feature from the dataset and the ratio between variances from two different samples are formulated [33]. For completion of the ANOVA technique, the below step is applied [33]:

Step 1: All features are selected from the original dataset

Step 2: The target feature function from scikit-learn is calculated using ANOVA F-score for each feature. Below (2), (3), (4) are the following formula to calculate ANOVA.

$$F = \frac{\text{variance between groups}}{\text{variance within groups}} \quad (2)$$

$$\text{Variance between groups} = \frac{\sum_{i=1}^n n_i (\bar{Y}_i - \bar{Y})^2}{(k-1)} \quad (3)$$

$$\text{Variance within groups} = \frac{\sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2}{(n-k)} \quad (4)$$

Step 3: The result from the test is used to perform feature selection which enables the removal of features that are unrelated to the target variable. The most influential features with the lowest variance are chosen in the experiment and tested with SelectKBest(); K represents the number of features for the final dataset.

Step 4: The number of features(n) with the highest ranking is used to create various feature subsets.

Research conducted by [37], shows the use of ANOVA can enhance the accuracy which is a 9.1% increase from 72.70%.

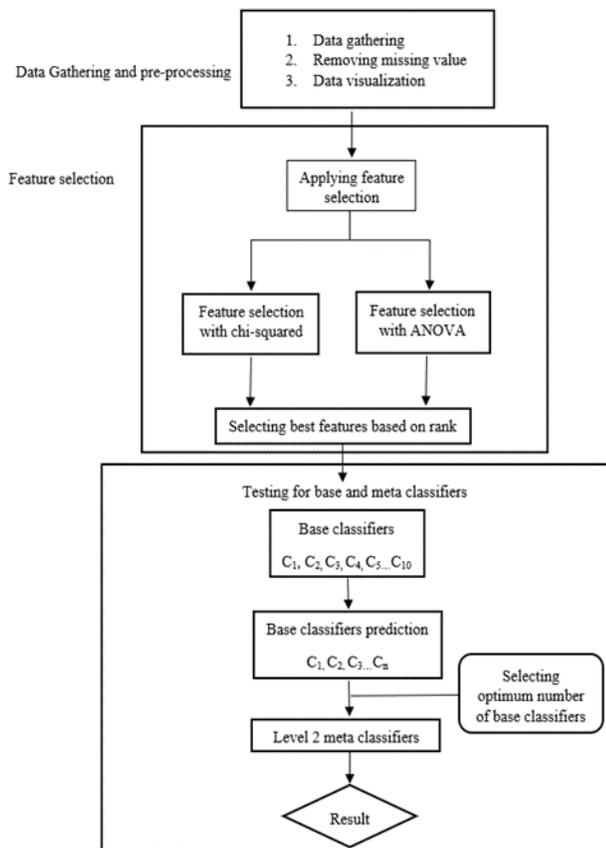


Fig. 3. Proposed framework for feature selection method

About Fig. 3, several steps are applied including data gathering and pre-processing before feature selection is applied. 13 attributes from the dataset are extracted to remove the missing value and visualize the data accordingly. Before we go deeper for base and meta classifiers, the feature selection method is applied to the data. Feature selection with the chi-squared technique is applied for categorical features while the ANOVA technique is applied for numerical features. Accuracy is tested for each feature selection method and the best accuracy is selected before we filter the important features. From the experiment made, five important features have been sorted out.

The data are then tested for base and meta-classifier methods. Ten base algorithms consisting of logistic regression (LR), support vector classifier (SVC), random forest (RF), extra tree classifier (ETC), naïve bayes (NB), extra gradient boosting (XGB), decision tree (DT), k-nearest neighbor (KNN), multilayer perceptron (MLP), and stochastic gradient descent (SGD) is tested and result obtained is used to find the optimum number of base classifiers. Then, meta-classifiers are applied for MLP, LR, NB, and SVC algorithms.

#### IV. RESULTS AND DISCUSSION

The proposed work is using the chi-squared method for categorical features and ANOVA for the numerical feature. 13 features from the UCI dataset are reduced to five features and tested accordingly. Highly rank of features are tested using the required method and there is an improvement in terms of accuracy for each algorithm.

Table III will further explain the involvement of five attributes for chi-squared and ANOVA and the achieved accuracy for each feature selection method.

TABLE III. FIVE CHOSEN ATTRIBUTES FROM CHI-SQUARED AND ANOVA

Selected attributes	Data Dictionary
$X_{12} = ca$	Number of major vessels
$X_3 = cp$	Chest pain type: Value 0= typical angina, Value 1=atypical angina, Value 2=non-anginal
$X_9 = exang$	exercise induced angine: 1=yes, 0=no
$X_{10} = oldpeak$	ST depression
$X_8 = thalach$	max heart rate achieved

Chi-squared and ANOVA technique feature selection was the focus of the subsequent testing phase. Ca, cp, exang, oldpeak, and thalach was chosen as the first five attributes selection which is superior to those of another feature selection algorithm.

Accuracy tests for both the base and meta classifiers using these five features and the result show an improvement from the accuracy of base classifiers. Results for both techniques of feature selection which are chi-squared and ANOVA are contracted in the following table.

TABLE IV. ACCURACY OF CLASSIFIERS BEFORE AND AFTER FEATURE SELECTION IS APPLIED

Algorithms	Feature selection					
	Base Classifier	Base classifier (after FS applied)	Meta classifier	Meta-classifiers (after FS applied)		
LR	85.24	91.80	90.16	93.44		
RF	83.60	86.89				
KNN	81.96	86.89				
DT	72.13	81.97				
NB	85.24	88.52				
SVC	86.88	91.80			83.60	91.80
XGB	85.24	81.97				
MLP	88.52	90.16			88.52	91.80
SGD	83.60	86.89				
ETC	86.88	83.61				

From Table IV, logistic regression obtains the highest accuracy compared to the other nine algorithms. For the level 1 base classifier, 85.24% is achieved before feature selection is applied and increases to 6.56% after feature selection is applied. Level 2 meta-classifier, increase from 90.16% to 93.44%.

For SVC, the accuracy increases by 4.92% from 86.88% for base classifiers and meta-classifiers, the accuracy achieved is 91.80% from 83.60%. MLP achieved 90.16% accuracy from 88.52% for base classifiers while the accuracy spike from 88.52% to 91.80% for meta-classifiers.

Classification and regression trees (CART) have an acquired accuracy of 87.65%, according to the literature [11]. The author makes an effort to boost precision by employing feature selection and an ensemble technique. There has been some improvement, but the accuracy is still low. In the current research, we suggested the same process but with a new set of features and an alternative method of feature selection. Logistic regression was able to provide a success rate of 93.44 percent, which is an increase over earlier efforts.

## V. CONCLUSION

The main goal of this work is to develop hybrid feature selection method for heart disease prediction that combines chi-squared and ANOVA approaches. ANOVA is used to choose numerical data, whereas Chi-squared is used to pick categorical features. The involved algorithms are logistic regression, k-nearest neighbor, decision tree, random forest, gaussian naive bayes, extra gradient boosting, support vector classifier, multilayer perceptron, stochastic gradient descent, and additional tree classifier. Various algorithms are tested for base classifiers. The meta-classifier is evaluated using the logistic regression, support vector classifier, and multilayer perceptron methods. Then, feature selection techniques are used to evaluate the base and meta-classifiers.

We decided to assess the efficacy of two distinct feature-selection algorithms in this study. Chi-squared tests and analysis of variance are utilized as feature selection methods. The experimental results show that the accuracy of heart disease prediction may be improved by employing the hybrid feature selection technique.

In addition to the utilization of feature selection techniques, the selected features from the dataset are also something that have to be emphasized. The chi-squared test and the analysis of variance (ANOVA) are used to evaluate the results of the experiment regarding five characteristics, namely ca, cp, exang, oldpeak, and thalach. The logistic regression method had a performance that was 93.44% better than the other ensemble stacking techniques. Because the accuracy of the approach might change depending on the dataset that is being used, it will be possible in the future to evaluate the technique of feature selection using a variety of different datasets.

## REFERENCES

- [1] M. Batta, "Machine Learning Algorithms - A Review," *Int. J. Sci. Res.*, vol. 18, no. 8, pp. 381–386, 2018, doi: 10.21275/ART20203995.
- [2] A. Mohd Ghazi, C. K. Teoh, and A. A. Abdul Rahim, "Patient profiles on outcomes in patients hospitalized for heart failure: a 10-year history of the Malaysian population," *ESC Hear. Fail.*, vol. 9, no. 4, pp. 2664–2675, 2022, doi: 10.1002/ehf2.13992.
- [3] N. A. F. Abu Bakar et al., "Association between a dietary pattern high in saturated fatty acids, dietary energy density, and sodium with coronary heart disease," *Sci. Rep.*, vol. 12, no. 1, pp. 1–11, 2022, doi: 10.1038/s41598-022-17388-5.
- [4] N. Endut, W. M. A. F. W. Hamzah, I. Ismail, M. K. Yusof, Y. A. Baker, and H. Yusoff, "A Systematic Literature Review on Multi-Label Classification based on Machine Learning Algorithms," *TEM J.*, vol. 11, no. 2, pp. 658–666, 2022, doi: 10.18421/TEM112-20.
- [5] I. Tougui, A. Jilbab, and J. El Mhamdi, "Heart disease classification using data mining tools and machine learning techniques," *Health Technol. (Berl.)*, vol. 10, no. 5, pp. 1137–1144, 2020, doi: 10.1007/s12553-020-00438-1.
- [6] S. I. Ansarullah, S. M. Saif, P. Kumar, and M. M. Kirmani, "Significance of Visible Non-Invasive Risk Attributes for the Initial Prediction of Heart Disease Using Different Machine Learning Techniques," *Comput. Intell. Neurosci.*, vol. 2022, 2022, doi: 10.1155/2022/9580896.
- [7] M. Hall and L. Smith, "Feature subset selection: a correlation based filter approach," *Proc. Int. Conf. Neural Inf. Process. Intell. Inf. Syst.*, pp. 855–858, 1998.
- [8] S. E. Fienberg, "The Use of Chi-Squared Statistics for Categorical Data Problems," *J. R. Stat. Soc. Ser. B*, vol. 41, no. 1, pp. 54–64, 1979, doi: 10.1111/j.2517-6161.1979.tb01057.x.
- [9] P. Schober and T. R. Vetter, "Chi-square Tests in Medical Research," *Int. Anesth. Res. Soc.*, vol. 129, no. 2, p. 2019, 2019.
- [10] M. S. Pathan, A. Nag, M. M. Pathan, and S. Dev, "Analyzing the impact of feature selection on the accuracy of heart disease prediction," *Healthc. Anal.*, vol. 2, p. 100060, 2022, doi: 10.1016/j.health.2022.100060.
- [11] S. Diwan, G. S. Thakur, S. K. Sahu, M. Sahu, and N. K. Swamy, "Predicting Heart Diseases through Feature Selection and Ensemble Classifiers," *J. Phys. Conf. Ser.*, vol. 2273, no. 1, 2022, doi: 10.1088/1742-6596/2273/1/012027.
- [12] S. Bashir, Z. S. Khan, F. Hassan Khan, A. Anjum, and K. Bashir, "Improving Heart Disease Prediction Using Feature Selection Approaches," *Proc. 2019 16th Int. Bhurban Conf. Appl. Sci. Technol. IBCAST 2019*, pp. 619–623, 2019, doi: 10.1109/IBCAST.2019.8667106.
- [13] A. Ismail et al., "Development of COVID-19 Health-Risk Assessment and Self-Evaluation (CHaSe): a health screening system for university students and staff during the movement control order (MCO)," *Netw. Model. Anal. Heal. Informatics Bioinforma.*, vol. 11, no. 1, 2022, doi: 10.1007/s13721-022-00357-3.
- [14] R. Spencer, F. Thabtah, N. Abdelhamid, and M. Thompson, "Exploring feature selection and classification methods for predicting heart disease," *Digit. Heal.*, vol. 6, pp. 1–10, 2020, doi: 10.1177/2055207620914777.
- [15] R. Mohd Rosdan, W. S. Wan Awang, and W. A. Wan Abu Bakar, "Comparison of affinity degree classification with four different classifiers in several data sets," *Int. J. Adv. Technol. Eng. Explor.*, vol. 8, no. 75, pp. 247–257, 2021, doi: 10.19101/IJATEE.2020.762106.
- [16] K. Dissanayake and M. G. M. Johar, "Comparative study on heart disease prediction using feature selection techniques on classification algorithms," *Appl. Comput. Intell. Soft Comput.*, vol. 2021, 2021, doi: 10.1155/2021/5581806.

- [17] A. Alfaidi, R. Aljuhani, B. Alshehri, H. Alwadei, and S. Sabbeh, "Machine Learning: Assisted Cardiovascular Diseases Diagnosis," *Int. J. Adv. Comput. Sci. Appl.*, vol. 13, no. 2, pp. 135–141, 2022, doi: 10.14569/IJACSA.2022.0130216.
- [18] A. Lakshmanarao, A. Srisaila, and T. S. R. Kiran, "Heart disease prediction using feature selection and ensemble learning techniques," *Proc. 3rd Int. Conf. Intell. Commun. Technol. Virtual Mob. Networks, ICICV 2021*, no. Icicv, pp. 994–998, 2021, doi: 10.1109/ICICV50876.2021.9388482.
- [19] R. Shanthakumari, C. Nalini, S. Vinothkumar, E. M. Roopadevi, and B. Govindaraj, "Multi Disease Prediction System using Random Forest Algorithm in Healthcare System," *2022 Int. Mob. Embed. Technol. Conf. MECON 2022*, pp. 242–247, 2022, doi: 10.1109/MECON53876.2022.9752432.
- [20] M. Jena and S. Dehuri, "Decision tree for classification and regression: A state-of-the art review," *Inform.*, vol. 44, no. 4, pp. 405–420, 2020, doi: 10.31449/INF.V44I4.3023.
- [21] P. A. Moreno-Sanchez, "Development of an Explainable Prediction Model of Heart Failure Survival by Using Ensemble Trees," *Proc. - 2020 IEEE Int. Conf. Big Data, Big Data 2020*, pp. 4902–4910, 2020, doi: 10.1109/BigData50022.2020.9378460.
- [22] M. Kavitha, G. Gnaneswar, R. Dinesh, Y. R. Sai, and R. S. Suraj, "Heart Disease Prediction using Hybrid machine Learning Model," *Proc. 6th Int. Conf. Inven. Comput. Technol. ICICT 2021*, pp. 1329–1333, 2021, doi: 10.1109/ICICT50816.2021.9358597.
- [23] R. Zebari, A. Abdulazeez, D. Zeebaree, D. Zebari, and J. Saeed, "A Comprehensive Review of Dimensionality Reduction Techniques for Feature Selection and Feature Extraction," *J. Appl. Sci. Technol. Trends*, vol. 1, no. 2, pp. 56–70, 2020, doi: 10.38094/jastt1224.
- [24] M. Kang and J. Tian, "Machine Learning: Data Pre-processing," *Progn. Heal. Manag. Electron.*, pp. 111–130, 2018, doi: 10.1002/9781119515326.ch5.
- [25] W. M. N. W. Z. Fakhithah Ridzuan, "A Review on Data Cleansing Methods for Big Data," *Elsevier*, p. 8, 2019. doi: 10.1016/j.procs.2019.11.177.
- [26] J. Jo, "Effectiveness of Normalization Pre-Processing of Big Data to the Machine Learning Performance," *J. KIECS.*, vol. 14, no. 3, pp. 547–552, 2019, doi: <https://doi.org/10.13067/JKIECS.2019.14.3.547>.
- [27] M. J. Hamid Mughal, "Data mining: Web data mining techniques, tools and algorithms: An overview," *Int. J. Adv. Comput. Sci. Appl.*, vol. 9, no. 6, pp. 208–215, 2018, doi: 10.14569/IJACSA.2018.090630.
- [28] H. Kaur, H. S. Pannu, and A. K. Malhi, "A systematic review on imbalanced data challenges in machine learning: Applications and solutions," *ACM Comput. Surv.*, vol. 52, no. 4, 2019, doi: 10.1145/3343440.
- [29] S. Sreelakshmi and K. G. Preetha, "Innovations in Bio-Inspired Computing and Applications," *Adv. Intell. Syst. Comput.*, vol. 424, no. Ibica, pp. 139–149, 2016, doi: 10.1007/978-3-319-28031-8.
- [30] S. Noelia, "Filter Methods for Feature Selection – A," pp. 178–187, 2007.
- [31] Y. B. Wah, N. Ibrahim, H. A. Hamid, S. Abdul-Rahman, and S. Fong, "Feature selection methods: Case of filter and wrapper approaches for maximising classification accuracy," *Pertanika J. Sci. Technol.*, vol. 26, no. 1, pp. 329–340, 2018.
- [32] S. E. Awan, M. Bennamoun, F. Sohel, F. M. Sanfilippo, B. J. Chow, and G. Dwivedi, "Feature selection and transformation by machine learning reduce variable numbers and improve prediction for heart failure readmission or death," *PLoS One*, vol. 14, no. 6, pp. 1–13, 2018, doi: 10.1371/journal.pone.0218760.
- [33] K. Dissanayake and M. G. M. Johar, "Comparative study on heart disease prediction using feature selection techniques on classification algorithms," *Appl. Comput. Intell. Soft Comput.*, vol. 2021, p. 17, 2021, doi: 10.1155/2021/5581806.
- [34] F. Kamalov, H. H. Leung, and S. Moussa, "Monotonicity of the  $\chi^2$  - statistic and Feature Selection," *Ann. Data Sci.*, vol. 9, no. 6, pp. 1223–1241, 2022, doi: 10.1007/s40745-020-00251-7.
- [35] N. Alotaibi and M. Alzahrani, "Comparative Analysis of Machine Learning Algorithms and Data Mining Techniques for Predicting the Existence of Heart Disease," *Int. J. Adv. Comput. Sci. Appl.*, vol. 13, no. 7, pp. 810–818, 2022, doi: 10.14569/IJACSA.2022.0130794.
- [36] R. L. Plackett, "Karl Pearson and the Chi-squared Test," *Int. Stat. Rev.*, vol. 51, no. 1, pp. 59–72, 1983.
- [37] M. F. Ihsan, S. Mandala, and M. Pramudyo, "Study of Feature Extraction Algorithms on Photoplethysmography (PPG) Signals to Detect Coronary Heart Disease," *2022 Int. Conf. Data Sci. Its Appl. ICoDSA 2022*, vol. 4, no. 2, pp. 300–304, 2022, doi: 10.1109/ICoDSA55874.2022.9862855.

# Automatic Extraction of Indonesian Stopwords

Harry Tursulistyo<sup>1</sup>, Yani Achsan<sup>1</sup>, Heru Suhartanto<sup>2</sup>, Wahyu Catur Wibowo<sup>3</sup>, Deshinta A. Dewi<sup>4</sup>, Khairul Ismed<sup>5</sup>

Faculty of Computer Science, University of Indonesia, Depok, Indonesia<sup>1,2,3</sup>  
Faculty of Science & Engineering, Universitas Paramadina, Jakarta, Indonesia<sup>1</sup>  
INTI International University & Colleges, Nilai, Malaysia<sup>4</sup>  
National Research and Innovation Agency of Republic of Indonesia, Indonesia<sup>5</sup>

**Abstract**—The rapid growth of the Indonesian language content on the Internet has drawn researchers' attention. By using natural language processing, they can extract high-value information from such content and documents. However, processing large and numerous documents is very time-consuming and computationally expensive. Reducing these computational costs requires attribute reduction by removing some common words or stopwords. This research aims to extract stopwords automatically from a large corpus, about seven million words, in the Indonesian language downloaded from the web. The problem is that Indonesian is a low-resource language, making it challenging to develop an automatic stopword extractor. The method used is Term Frequency – Inverse Document Frequency (TF-IDF) and presents a methodology for ranking stopwords using TFs and IDF, which is applicable to even a small corpus (as low as one document). It is an automatic method that can be applied to many different languages with no prior linguistic knowledge required. There are two novelties or contributions in this method: it can show all words found in all documents, and it has an automatic cut-off technique for selecting the top rank of stopwords candidates in the Indonesian language, overcoming one of the most challenging aspects of stopwords extraction.

**Keywords**—Stopwords extraction; attributes reduction; TF-IDF; large corpus; Indonesian stopwords; NLP

## I. INTRODUCTION

Stopwords are any common words that carry low information content [1]. Despite their high occurrence, they only add a little semantic data to the document [2]. They are also referred to as negative dictionary or noise words. They cause a small retrieval degree and prediction outcomes. Since they make up a considerable portion of the documents, text-mining tasks will be very computationally intensive. This high computational cost is caused by the dimensionality curse and requires larger computer memory and computational time. Furthermore, in information retrieval experiments, it has been shown that removing stopwords improves precision significantly when compared with when they are not removed [3, 4]. Stopwords also play a significant role in feature extraction [5, 6], topic modeling [7], classification [8], ontology construction [9], and keyword extraction [10].

There are two categories of stopwords: domain-specific and general. Domain-specific stopwords are a set of words that make no discriminant value inside a specific context or domain. They differ from one domain to another domain. For example, the word “learning” could be a stopword in the education domain, or the word “machine” could be a stopword

in the machinery domain, but neither of those words is a stopword in the computer science domain. On the other hand, general stopwords are a list of stopwords or stoplists that are not specific to one domain and are usually available to download as a public domain object.

General stopwords are the most used in Natural Language Processing (NLP) because of their availability, and it takes a considerable effort to develop a domain-specific stoplist. It is easier to create a domain-specific stoplist based on a general stoplist by adding and/or removing some terms. General stoplists, however, need to be updated frequently. In addition, over time, the use of some ordinary words has altered subjects on social aspects such as industrial revolution changes, new media, cultural shifts, and education. For these reasons, reviewing, updating, and adjusting existing stoplists is essential [5]. Updating a general stoplist can be done manually, but it takes time and may omit the latest stopwords. This problem can be solved automatically by developing a general stoplist.

Researchers have developed many methods for automatically creating stoplists, especially in English, since decades ago. Since then, many methods have been developed to create English stoplists. In contrast, there are relatively few studies to develop a stoplist for non-English languages like Indonesian. The problem is that Indonesian is a low-resource language, making it challenging to develop an automatic stopword extractor.

There were only two research documents about general stopwords extraction in Indonesian [11, 12]. Those documents show 394 and 330 general stopwords extracted from Kompas daily newspaper. Both of stopwords lists extracted using Term Frequency (TF) method, it is a rare method to use in extracting stopwords. Most researchers use a combination of Term Frequency and Inverse Document Frequency (TF-IDF) in NLP. Unfortunately, TF cannot detect words that occur in all documents and cannot implement threshold to limit the number of generated stopwords.

This research paper aims to solve the problem above and develop an up-to-date general stoplist in the Indonesian language. The method used is crawling recent news from an Indonesian online newspaper's website to gather data and make the required dataset. The stopwords extraction method uses TF-IDF.

This document is organized in this way; the following section comprises a brief coverage of the present literature in the areas of automatic stopwords extraction, the methods used for stopwords extraction, and the experiments. Then we

describe the results of this research. We conclude this research document by presenting the advancement of our methods compared to previous works.

## II. RELATED WORKS

Many methods have been used to develop stoplists. Some of them are frequency-based approach [13], Bidirectional Long Short term memory (BiLSTM) [14], Word Embedding [15], Finite Automata [16], and utilizing characteristic and discriminant analysis [17]. The dataset or corpus used to extract or identify stopwords vary. Some researchers used corpus from an online newspaper [18], social network [19], or patent [20]. As the purpose of this research is to develop an Indonesian general stoplist, only relevant research papers are discussed.

There are three research papers discussing the development of the Indonesian stoplist. One of them only involves developing a cuisine-specific stoplist for the Indonesian language [12]. However, since this research aims to develop a general Indonesian stoplist, only the other two papers are reviewed further here based on their proposed method.

Fadillah Z Tala, in his master thesis [11], proposed an Indonesian stoplist because there was no Indonesian stoplist that could be used in his experiment in information retrieval. In his work, he created the dataset based on articles from the “Kompas daily” newspaper. He downloaded the articles every day for one year long, starting from the beginning of January 2001 until the end of December 2001. The total number of articles was 3160. The result of his experiment was 394 stopwords in Bahasa Indonesia.

Yudi Wibisono has created a stoplist in his coursework [12]. As the source of his dataset, he also used articles from the “Kompas daily” newspaper. He used several hundred articles to create 330 stopwords. The method used was Term Frequency, like Tala’s work, but he removed some words manually.

Tala and Wibisono used the Term Frequency (TF) method to extract stopwords in their work. These days, Term Frequency-Inverse Document Frequency (TF-IDF) is another method that is generally used in information retrieval systems. TF-IDF is one of the traditional methods based on statistics [21]. It has been used in many different applications, such as document clustering [22], text classification [23], detection of domain name generation algorithms [24], and comparing research trends [25]. Term frequency or word frequency is a rarer method used in information retrieval systems compared to TF-IDF.

## III. METHODS

Different methods were used in each stage of this research. As shown in Fig. 1, the steps for this study were differentiated into three stages: data gathering, pre-processing or data cleaning, and stopwords extraction.

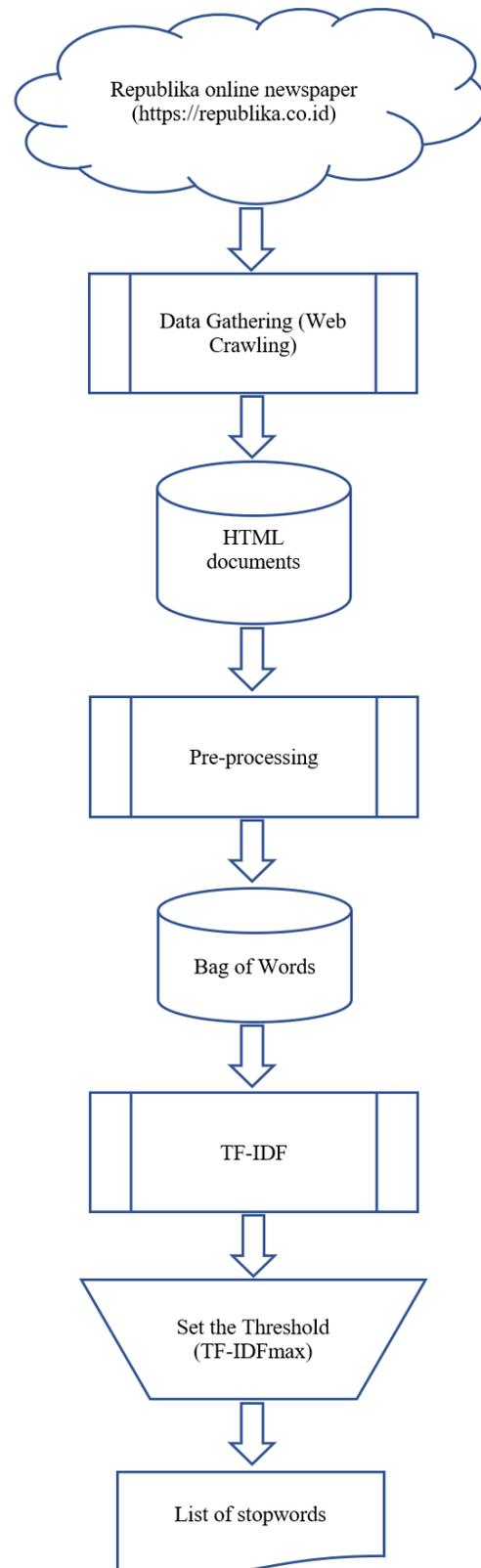


Fig. 1. Steps for this study

This work is different from the previous studies in some stages. First, the data source in the data gathering stage of this research is crawled from the “Republika daily” newspaper, whereas the previous studies used data from the “Kompas daily” newspaper. Moreover, in their studies, they used Term Frequently (TF), but in this work, we used the TF-IDF method, a combination of the Term Frequently and Inverse Document Frequency methods.

#### A. Data Gathering

The dataset or corpus for this research was gathered from Republika, an Indonesian online newspaper. The method used to gather the data was “Focused Web Crawling” [26, 27]. “Focused Web Crawling” is a method to download or harvest particular data from websites, commonly from one website. The crawler was developed using Python programming language to crawl web addresses from the Republika website. There were 6111 articles downloaded, containing 6,947,178 words, 87,998 of which were unique.

#### B. Pre-Processing

Pre-processing is a required process to clean the dataset. Some steps in pre-processing are case folding, HTML tags removal, special characters removal, tokenizing, dealing with missing data, dealing with data error, and stemming. These stopwords extractions implement pre-processes are as follows:

- Case folding: Converting characters from uppercase to lowercase. The fastest and simplest way is entirely changing words to lowercase, including words occurring in a sub-title or title and words at the beginning of a sentence. Since some papers used uppercase for *Term Frequency* and others used lowercase *term frequency*, so in our research, we converted all words into lowercase, which means that we treat those two phrases as the same phrase.
- HTML tags removal: Removing all HTML tags, scripts, and other metadata from HTML documents is mandatory. It returns clean texts from documents in HTML format.
- Special characters removal: It includes removing punctuations, numbers, and other non-text characters. Examples of the special characters that removed from the text are @%&+?,:;-'()[]{}|`~"0123456789.
- Tokenizing: It separates each word from documents into an array of items or a bag of words.

#### C. Stopwords Extraction

Extracting stopwords from Indonesian documents is the primary purpose of this study. The stopwords extraction from the dataset used the TF-IDF method after the pre-processing steps. Eq. (1)-(5) present this TF-IDF:

$$TF - IDF(\omega_i) = tf(\omega_i) \times idf(\omega_i) \quad (1)$$

$$tf(\omega_i) = \frac{n_{ij}}{\sum_{k=1}^m n_{kj}} \quad (2)$$

$$idf(\omega_i) = \log\left(\frac{N}{df(\omega_i)}\right) \quad (3)$$

$$df(\omega_i) = |\{j: \omega_i \in d_j\}| \quad (4)$$

where  $f(\omega_i)$  is frequency of occurrence of term or word  $\omega_i$  in document  $j$ , and  $N$  is total number of all documents in document collection  $\{d_j\}$ .  $df(\omega_i)$  indicates the number of documents contain term  $\omega_i$  in the document collection.  $n_{ij}$  is the number of occurrences of  $i$ th term appearing in  $j$ th document.  $n_{kj}$  is occurrence frequency of  $k$ th term appearing in  $j$ th document.  $|\{j: \omega_i \in d_j\}|$  is number of document consisting  $i$ th term. Getting the value of each term in every document is done by examining every term in the document collection or corpus.

For the whole document collection, corpus or dataset, the average of TF,  $tf(\omega_i)$ , is divided by the number of documents consisting of term  $\omega_i$ . Thus, the TF-IDF formula of term  $\omega_i$  for the whole document collection is:

$$TF - IDF(\omega_i) = \frac{\sum_j^N tf(\omega_{ij})}{df(\omega_i)} \times idf(\omega_i) \quad (5)$$

## IV. EXPERIMENTS

Several experiments have been done to find the methods. For example, the data gathering method should be tried many times before we can harvest the data automatically. It is because the articles or documents are spread into tens of categories or sub-categories in the data source (<https://www.republika.co.id/>), such as News, Playing Games, Economics, Football, Islam Digest, or International. Since the structure of these web pages was not crawler friendly, we used Focused Web Crawling strategy to handle them. The data is then processed using the discussed pre-processing methods.

The pre-processing methods used were standard methods for Natural Language Processing. Our experiments regarding pre-processing proceeded smoothly. All pre-processes were done automatically by using applications developed in Python languages. The Python language was chosen because of its many machine learning libraries, especially for NLP. Later, a bag of words or an array of terms resulted from pre-processes fed into the TF-IDF method.

We used the TF-IDF formula shown in (5) to extract stopwords. Since there is no need for a training dataset, this NLP approach is categorized as an unsupervised machine learning method. It contains  $idf(\omega_i)$  that comes from equation (3). It normalizes equation (5), limiting results of  $TF - IDF(\omega_i)$  between zero and one.

If  $TF - IDF(\omega_i)$  is equal to 0, it means that the  $i$ -th word ( $\omega_i$ ) exists in all documents. Table I shows three words contained in all documents that are *republikacoid* (republika.co.id), *wib* (west Indonesian time zone), and *lainnya* (others). The greater value of  $TF - IDF(\omega_i)$  denote that the word is less significant to be a stopwords. Fig. 2 shows the correlation between  $TF - IDF_{\max}$  and the number of stopwords extracted in the logarithmic scale. This figure shows that those words are stopwords if maximum of  $TF - IDF(\omega_i)$  is equal to 1.

TABLE I. SAMPLE OF EXTRACTED STOPWORDS USING TWO DIFFERENT METHODS

No	Tala's & Wibisono's Method			This Research Method		
	Rank	Term	Frequency	Rank	Term	TF-IDF
1	1	yang	28,913	1	republikacoid	0
2	2	dan	27,074	2	wib	0
3	3	di	25,634	3	lainnya	0
4	4	untuk	13,760	4	terpopuler	0.0000007
5	5	dari	12,241	5	foto	0.0000007
6	6	dengan	11,839	6	terkait	0.0000011
7	7	pada	10,627	7	di	0.0000022
8	8	ini	10,478	8	home	0.0000039
9	9	photo	9,547	9	republikaid	0.0000044
10	10	dalam	8,525	10	copyright	0.0000045
11	17	republikacoid	6,116	11	reserved	0.0000045
12	68	newsroom@rolrepublikacoid	2,759	12	right	0.0000046
13	69	sekretariat@republikacoid	2,759	13	all	0.0000047
14	70	update	2,743	52	sekretariat@republikacoid	0.0002252
15	71	memicu	2,737	56	newsroom@rolrepublikacoid	0.0002282
16	72	marketing@republikacoid	2,734	57	marketing@republikacoid	0.0002282
17	73	kepada	2,727	58	us	0.0002283
18	74	direncanakan	2,703	59	gerai	0.0002289
19	75	tergolong	2,680	60	about	0.0002289
20	76	jis	2,674	61	copy	0.0002296

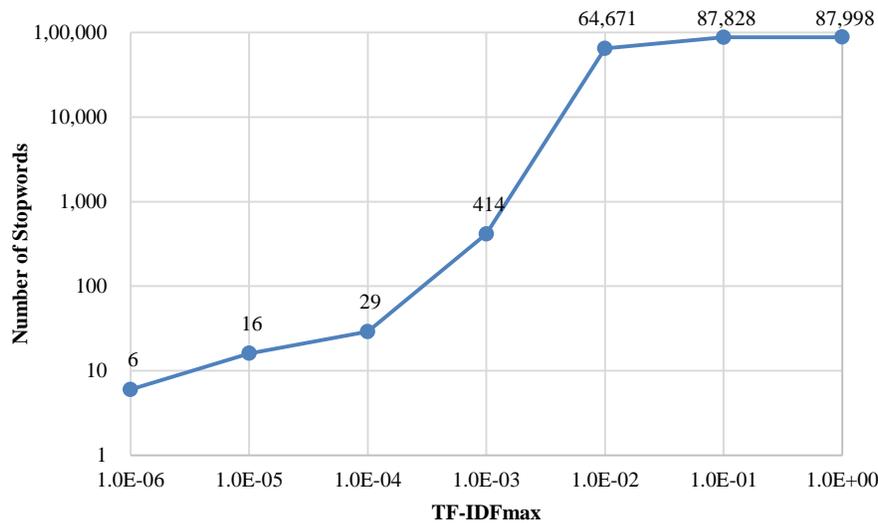


Fig. 2. The correlation of the threshold (TF-IDFmax) and the number of stopwords extracted in the logarithmic scale

## VI. RESULTS

The number of extracted stopwords using the proposed method depends on the defined TF-IDF threshold. For example, the system extracted only six stopwords for the maximum of TF-IDF  $10^{-6}$ , and 414 stopwords if the maximum of TF-IDF increased to  $10^{-3}$ . However, for TF-IDF<sub>max</sub> equal to 0.01, the number of stopwords is blown up to 64,571. The cut-off of the number of stopwords can be done by setting the value of TF-IDF<sub>max</sub>. Since the range of TF-IDF is 0 to 1, the threshold can be maintained constantly. For example, if the threshold is set to 0.001 and the number of documents doubled the number of stopwords generated by TF-IDF does not change significantly. If we only use TF to extract stopwords and set the threshold to 8,000 and double the number of documents, then the frequency of stopwords generated might be doubled, resulting in the number of stopwords changing significantly as shown in Table I under Tala's and Wibisono's method.

The extracted stoplist contains some words specific to the dataset. For example, since the dataset or document collection source is *Republika* online daily newspaper, then there are some words with TF-IDF equal to zero. It means that those words occur in all documents. Table I shows the sample of extracted keywords from the same document collection using two different methods. Results in the left column are based on the previous researcher's method, and the right column is based on the method proposed in this work. As shown in this table, other methods can not reveal words that occurred in all documents.

Analyzing the top 100 extracted stopwords shows that the method used in this research, TF-IDF, is better than the previous methods. First, this research output can reveal the words that occur in all documents and place it in the top ranks, while the old method can reveal only two words and place them in ranks 17 and 26. Second, TF-IDF method can expose all words in the sentence "copyright ... all right reserved" that occur in most of the documents, where the old method cannot reveal any of those words.

## VII. CONCLUSIONS AND FUTURE WORKS

Stopwords extraction using TF-IDF has three advances compared to TF. First, it can detect words that occur in all documents with TF-IDF equal to zero. Second, it can implement threshold to limit the number of generated stopwords. Third, it can expose all words that occur in most of the documents and place it in the top ranks.

This research can be improved for future works in two ways. First, the documents in the corpus should be classified by its domain because stopwords for one domain are different from other domains. Secondly, develop a recommender system, a web-based application for the stopwords extraction that can be accessed by public.

## ACKNOWLEDGMENT

Excellent Research Grants of Higher Education (PT-UPT), Directorate General of Higher Education, Research, and Technology, Ministry of Education, Culture, Research, and Technology of the Republic of Indonesia (Ditjen Dikti -

Kemdikbud) titled "Representation of Multi Talents of Covid-19 Expert Based on indexed publication data, 2019 & 2020".

## REFERENCES

- [1] J. Kaur and P. K. Buttar, "A systematic review on stopword removal algorithms," *International Journal on Future Revolution in Computer Science & Communication Engineering*, vol. 4, no. 4, pp. 207-210, 2018.
- [2] M. Dehghani and M. Manthouri, "Semi-automatic detection of Persian stopwords using FastText library," in 9781665402088, 2021.
- [3] S. Sahu and S. Pal, "Effect of stopwords in Indian language IR," *Sadhana - Academy Proceedings in Engineering Sciences*, vol. 47, no. 1, pp. -, 2022.
- [4] A. Bichi, R. Samsudin and R. Hassan, "Automatic construction of generic stop words list for hausa text," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 5, no. 3, pp. 1501-1507, 2022.
- [5] R. Arlitt, S. Khan and L. Blessing, "Feature engineering for design thinking assessment," in *International Conference on Engineering Design*, 2019.
- [6] K. Goucher-Lambert and J. Cagan, "Crowdsourcing inspiration: using crowd generated inspirational stimuli to support designer ideation," *Design Studies*, vol. 61, pp. 1-29, 2019.
- [7] H. Song, J. Evans and K. Fu, "An exploration-based approach to computationally supported design-by-analogy using D3," *AI EDAM*, vol. 34, pp. 444-457, 2020.
- [8] S. Urologin, "Sentiment analysis, visualization and classification of summarized news articles: a novel approach," (IJACSA) *International Journal of Advanced Computer Science and Applications*, vol. 9, no. 8, pp. 616-625, 2018.
- [9] F. Shi, L. Chen, J. Han and P. Childs, "A data-driven text mining and semantic network analysis for design information retrieval," *Journal of Mechanical Design*, vol. 139, no. 11, 2017.
- [10] B. Guda, B. K. Nuhu, J. Agajo and I. Aliyu, "Performance evaluation of keyword extraction techniques and stop word lists on speech-to-text corpus," *International Arab Journal of Information Technology*, vol. 20, no. 1, pp. 134-140, 2023.
- [11] F. Z. Tala, "A study of stemming effects on information retrieval in bahasa Indonesia," *Institute for Logic, Language and Computation Universiteit van Amsterdam The Netherlands, Amsterdam, The Netherlands*, 2003.
- [12] S. Wibisono and M. Utomo, "Dynamic stoplist generator from traditional Indonesian cuisine with statistical approach," *Journal of Theoretical and Applied Information Technology*, vol. 87, no. 1, pp. 92-98, 2016.
- [13] S. Gandotra and B. Arora, "Automated stop-word list generation for dogri corpus," *International Journal of Advanced Science and Technology*, vol. 28, no. 19, pp. 884-889, 2019.
- [14] K. Gorro, M. Ali, L. Lawas and A. Ilano, "Stop words detection using a long short term memory recurrent neural network," *ACM International Conference Proceeding Series*, pp. 199-202, 2021.
- [15] Z. Nassr, N. Sael and F. Benabbou, "Generate a list of stop words in Moroccan dialect from social network data using word embedding," *Ensa Marrakech;Faculte des Sciences et Techniques;Marrakech;Universite Abdelmalek Essaadi;Universite Cadi Ayyad*, 2021.
- [16] T. Kochhar and G. Goyal, "Design and implementation of stop words removal method for Punjabi language using finite automata," *Lecture Notes on Data Engineering and Communications Technologies*, vol. 106, pp. 89-98, 2022.
- [17] G. Armano, F. Fanni and A. Giuliani, "Stopwords identification by means of characteristic and discriminant analysis," in 9789897580741, 2015.
- [18] S. Gunasekara and P. Haddela, "Context aware stopwords for Sinhala text classification," in 9781538691366, 2018.
- [19] Y. Nezu and T. Miura, "Extracting stopwords on social network service," in 9781643680446, 2019.
- [20] S. Sarica and J. Luo, "Stopwords in technical language processing," *PLoS ONE*, vol. 16, no. 8 August, pp. -, 2021.
- [21] F. Lan, "Research on text similarity measurement hybrid algorithm with term semantic information and TF-IDF method," *Advances in Multimedia*, vol. 2022, pp. -, 2022.

- [22] J. ZHU, S. HUANG, Y. SHI, K. WU and Y. WANG, "A Method of k-means clustering based on TF-IDF for software requirements documents written in Chinese language," *IEICE Transactions on Information and Systems*, vol. 105, no. 4, pp. 736-754, 2022.
- [23] L. Xiang, "Application of an improved TF-IDF method in literary text classification," *Advances in Multimedia*, vol. 2022, pp. -, 2022.
- [24] H. Vranken and H. Alizadeh, "Detection of DGA-generated domain names with TF-IDF," *Electronics (Switzerland)*, vol. 11, no. 3, pp. -, 2022.
- [25] H. Toosi, M. Ghaaderi and Z. Shokrani, "Comparative study of academic research on project management in Iran and the world with text mining approach and TF-IDF method," *Engineering, Construction and Architectural Management*, vol. 29, no. 3, pp. 1553-1583, 2022.
- [26] J. Liu, X. Li, Q. Zhang and G. Zhong, "A novel focused crawler combining web space evolution and domain ontology," *Knowledge-Based Systems*, vol. 243, pp. -, 2022.
- [27] S. Rajiv and C. Navaneethan, "Hybrid gradient strategies in event focused web crawling," in 9781607685395, 2022.

# Software Effort Estimation through Ensembling of Base Models in Machine Learning using a Voting Estimator

Beesetti Kiran Kumar<sup>1</sup>, Saurabh Bilgaiyan<sup>2</sup>, Bhabani Shankar Prasad Mishra<sup>3</sup>

PhD Scholar, KIITs Deemed to be University, Bhubaneswar, Odisha, India<sup>1</sup>

Assistant Professor, Department of Information Technology, ANITs, Vishakhapatnam, India<sup>1</sup>

Assistant Professor, School of Computer Engineering, KIITs Deemed to be University, Bhubaneswar, Odisha, India<sup>2</sup>

Professor, School of Computer Engineering, KIITs Deemed to be University, Bhubaneswar, Odisha, India<sup>3</sup>

**Abstract**—For a long time, researchers have been working to predict the effort of software development with the help of various machine learning algorithms. These algorithms are known for better understanding the underlying facts inside the data and improving the prediction rate than conventional approaches such as line of code and functional point approaches. According to no free lunch theory, there is no single algorithm which gives better predictions on all the datasets. To remove this bias our work aims to provide a better model for software effort estimation and thereby reduce the distance between the actual and predicted effort for future projects. The authors proposed an ensembling of regressor models using voting estimator for better predictions to reduce the error rate to over the biasness provide by single machine learning algorithm. The results obtained show that the ensemble models were better than those from the single models used on different datasets.

**Keywords**—Machine learning; software effort estimation; voting; regression; evolutionary algorithms

## I. INTRODUCTION

For a given project, the effort estimation of software is always a burdensome task. For an extended period, team and finance managers strive to precisely calculate the effort, cost, and time while helping evaluate the project's schedule and budget parameters [1]. It is very tough to predict those specifications during the early stages of the project, where the scope of every module has yet to be marked, and when we still have no conclusive evidence for the ultimate functional requirements of the product [2]. Most frequently, insufficient knowledge of the affecting factors and the possible risks that can happen, or the work deadline fears, and conventional effort estimation methods [3], which are widely accepted by the opinions of various software domain experts, may sometimes lead to erroneous estimates. Because of these, the software product may not be delivered in time with the expected non-functional requirements. Though there are frequent improvements in the up-gradation of software development standards, surveys [4] show that only a quarter of the total number of beginnings is successful. These issues, which result in going beyond the budget or schedule, may lead to its termination [5]. Though the usage of agile methodology [6] reduced some concerns, project omissions are still occurring because of not having access to all the country's projects. For a country that just has a limit to its country's

projects, obtaining success in the projects is still a problem. All the managers of the project are claiming otherwise. The causes are the poor skillsets of the teams on the project and less bonding with stakeholders. There are high chances of more project success if the effort is well predicted from the beginning. But once the project's parameters are set, it's not a good idea to increase either the budget or the schedule because that could lead to risks that are hard to predict.

The client's approval for that scenario must be independent of the chosen process model for project development, time, and cost determination. For this case, some simple and easy conventional techniques that experts accept, such as PERT, CPM, etc., are primarily deployed [7]. Researchers started working on methods depending on software lines of code, and functional points as the previously mentioned techniques are vulnerable [8]. In various ways, the software parameters try to be up to date with the improved technologies. However, those techniques struggle to keep pace in the fastest-growing world, specifically with the reusability components and software dependencies that have already gone so far in their enhancement.

After considering all these drawbacks, researchers dug deep for efficient predictive techniques for effort, especially in data science areas [9]. This area is highly trusted, proving its potential with uncertain and unstructured data. Hence, it is believed that they can find the effort and duration way better than existing models. In other methods, they consider patterns in the previous data and do not rely on human influences, making them unique in their work. The factors behind this are systematic research that builds the best model for prediction to reduce the error rate for data. Biased models have been generated for some time. Their work is limited to a particular dataset, repeatedly underfitting or overfitting using varied ensemble techniques. The critical role is preparing data that has a crucial impact on the model, but divergent methods are to be used. Even the individual algorithms may not give an improved score, which is evident from various journals. The same can be repeated with effort and cost prediction when working with data and building models with those algorithms [10].

Though using all project parameters produced better results in the literature, some works included proper feature

selection strategies, eliminating irrelevant features for less computation usage and creating a unique effect with fewer features than the others. Some of them are the Genetic Algorithm [11], PSO [12], and WCO [13].

This work aims to improve the existing models on reliable machine learning algorithms for the best effort prediction. An averaging ensemble of various regressors proposes a hybrid model for this aim. Also, the proposed work conducted experiments on various datasets such as cocomo81, china, Desharnais, Maxwell, Kemerer, and Albrecht etc. to evaluate the behavior of the model with varying datasets.

## II. RELATED WORK

Amini et al. [14], in their paper combined two techniques, namely embedded and wrapper methods. The main motto of their writing is to integrate GA into regularized learning to improve prediction accuracy in regression problems. The outcome of their study reduced the dimension feature space by over 80% without affecting the accuracy. De Carvalho et al. [15] proposed an Extreme learning machine for forecasting software efforts. For selecting the best features, the Pearson correlation coefficient is used for feature selection. Extreme Learning Machines (ELM) are used with different numbers of hidden layers in their work. The ELM model values are compared with the models mentioned in the literature, namely LR, SVM, KNN, and MLP. The metric evaluated for comparison is MAE.

Ghosh et al. [16] proposed the binary variant of SFO for selecting features. This work compared ten state-of-the-art techniques and declared that BSFO based on adaptive hill climbing had shown better reliability. Carbonera et al. [17] surveyed over 120 studies and indicated that this study encouraged the researchers to minimize the space in the effort estimation. Chhabra et al. [18] worked in soft computing Fuzzy model along with PSO. This Fuzzy logic improved the existing COCOMO technique. The metric followed is MRE for result comparison. Ghatasheh et al. [19] proposed a firefly algorithm to optimize software effort. The results were better than the conventional models used earlier.

Wani et al. [20] worked on ANN and PSO. The limitation in their work is that the combination is giving better results for only the cocomo81 dataset. The ANN showed fast training speed than MLP. This method ended with lesser MdmRE and MRE than other models. Ali et al. [21] used all bio-inspired feature selection strategies with Support Vector and Random Forest regressors. The evaluation metrics considered are Correlation Coefficient (CC), Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Relative Absolute Error (RAE), and Root Relative Squared Error (RRSP).

Kodmelwar et al. [22] modified a neural network for effort prediction. The proposed method uses java for the front-end tool. The metrics used for comparison are PRED, MRE, MAE, MMRE, and RE. Desai et al. [23] combined ANN with cuckoo optimization, and experiments were performed on various datasets. This hybrid combination worked well then

all other literature models except for the ACO technique. Langsari et al. [24] optimized the parameters of the COCOMO II model using particle swarm optimization. PSO is a valuable strategy for resolving dataset uncertainty and optimizing the values. The author worked on the Turkish software industry dataset.

Hosni et al. [25] concentrated on parameter tuning ensemble using grid search optimization. The authors evaluated results over seven datasets to compare statistical measures, namely mean, median, and inverse ranked weighted mean. Used three algorithms, GS, PSO, and UC-Weka, and concluded that PSO gained over other approaches. Goyal et al. [26] proposed an SG5 neural network model trained on the Cocomo dataset and tested in the Kemerer dataset. It excelled over the traditional models. Padhy et al. [27] developed an Aging and survivability-related reusability optimization model, and the software metric estimation is done with the help of UML or Class diagrams. To overcome the limitations of ANN, some different Evolutionary Computing (EC) algorithms like Genetic Algorithms, Differential Evolution, and Particle Swarm Optimization (PSO) have been proposed. By implementing the above algorithms, the regression outputs are improved so that the results are significantly accurate and most effective.

Pospieszny et al. [28] proposed ensemble averaging with a 3-fold validation, namely SVM, MLP, and GLM, to predict both effort and duration. Here used the standard ISBSG dataset and considered the MMRE and PRED metrics. In their paper, Shekhar et al. [29] discussed various software cost estimation techniques and models. The authors classified these techniques into algorithmic and non-algorithmic, which helps the software team rule out the weaker methods and provides specific areas for considering an approach.

Venkatesh et al. [30] calculated the workforce to determine the cost and effort of the project, which outperformed other models, like regression models and neural networks. This work applied to several PROMISE datasets by considering RMSE as the root metric. Nassif et al. [31] worked on four different neural networks, the oldest projects used for training and the newest projects used for testing. Here ten-fold cross-validation is achieved. The author concluded that in 60% of datasets, CCNN performed better than other models, and on 40% of datasets, RBFNN performed better than others. Miandoab et al. [32] proposed a hybrid Algorithm using a particle swarm optimization algorithm and fuzzy logic.

Dizaji et al. [33] combined Ant Colony Optimization (ACO) and Lorentz transformation as Chaos Optimization Algorithm (COA). The meta-heuristic algorithms like ACO and COA are used to estimate the cost of the software. Mean Absolute Relative Error (MARE) is taken into consideration. Here the dataset is classified and distributed among the ACO and hybrid ACO and COA algorithms according to their functionalities. The results show that the performance is improved and efficient when the ACO algorithm is combined with COA.

### III. METHOD

The proposed approach introduces a novel method of using ensemble techniques with voting for software development effort estimation. This approach combines the strengths of multiple models and leverages the diversity of their predictions to improve accuracy. By investigating the impact of different factors on the accuracy of the ensemble with voting, this approach can provide insights into how to optimize the performance of the ensemble for different datasets and problems. The proposed approach can also have practical applications for software development organizations, as it can help them to make more accurate and informed decisions about project planning and resource allocation. The proposed architecture is illustrated in Fig. 1.

- Collect historical project data: Gather historical project data including information on the size of the project, the number of developers involved, the complexity of the software, and the amount of time and resources required to complete the project.
- Preprocess the data: Preprocess the data to remove any outliers or errors, and to convert the data into a format that can be used by the ensemble models.
- Train multiple estimation models: Train multiple estimation models on the preprocessed data, such as linear regression, decision trees, neural networks, and support vector machines.
- Implement the voting algorithm: Implement the voting algorithm to combine the predictions from the multiple models. There are different types of voting algorithms such as majority voting, weighted voting, and threshold voting.
- Evaluate the ensemble with voting: Evaluate the accuracy of the ensemble with voting using a validation set of historical data that was not used during training. Compare the performance of the ensemble with voting against individual models and other ensemble techniques.
- Investigate the impact of different factors: Investigate the impact of different factors on the accuracy of the ensemble with voting, such as the number of models, the type of models, the voting algorithm, and the size and quality of the historical data.
- Apply the ensemble with voting to new data: Apply the ensemble with voting to new software development projects to assess its accuracy and reliability in real-world scenarios.

#### A. Data Preprocessing

Preprocessing of data involves data cleansing approaches. It has a clear positive impact on training the machine learning models. It reduces the dataset's noise by filling in missing values, removing duplicate records, dropping unnecessary columns, etc. Finally, it produces the data in its best representation to be used for model building. Without preprocessing, models might learn the noise as an underlying pattern, leading to overfitting or underfitting the data. Here,

we dropped some attributes in our work, such as project ids, dates of projects, other categorical details, etc. We ignored the missing data records.

#### B. Normalization

Normalization is done as a second step, and it is essential to scale the features within a range for the model's performance. This normalization sets the feature scale from 0 to 1 and is implemented using the MinMax scalar in Python. In our datasets, we normalized all the input and output features.

$$x_{new} = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (1)$$

#### C. 5 - Cross Fold Validation

Cross-fold validation is an interesting technique, which makes our model more reliable. Instead of considering a particular subset for training and the remaining part for testing, it uses the entire dataset for training and testing purposes. A five-fold validation usually splits the entire dataset into five equal sets or folds, where for every time, four sets are used for training, and one set is used for testing. This process is repeated for four (k-1) iterations, i.e., all possible combinations, and it will give the average score of all iterations.

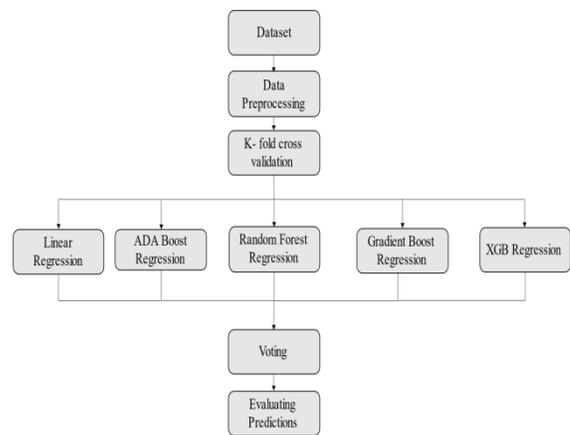


Fig. 1. Proposed architecture

#### D. Algorithms

In our work, we build a hybrid model with the help of five machine learning Regression algorithms. Each algorithm has a different structure in its implementation.

1) *Linear regression*: Linear Regression frames an equation for the given attributes to fetch the target variable. It assumes a linear relationship between the characteristics of a dataset. The equation is  $y = f(x)$ , where  $y$  represents the output variable and  $x$  is the set of input attributes. This algorithm performs better than complex models when the dataset is linear.

2) *Random forest*: Random Forest is a bagging model. It constructs several trees for prediction. Every tree is constructed from a subset of the training data. Every tree will give some effort for a test set. All predictions are averaged to

get the final estimate of how much work needs to be done, lowering the result's error rate.

3) *Boosting techniques*: Every boosting algorithm has a base model. After each iteration, a new weak learner is added to the sequence of learners; every iteration model reduces the residual effort. We implemented three boosting models in our work: Ada Boost, Gradient Boost (GB), and Extreme Gradient Boost. AdaBoost handles missing data well and undergoes no overfitting. It has fewer parameters to tune when needed and is sensitive to outliers. Gradient Boosting is a sequence of tree learners robust to outliers, depending on residuals. XGB has been showing better results than GB as it includes the calculation of similarity weights.

#### E. Voting

Every algorithm is unique in its background processing of data. Hence, all algorithms can find their patterns of data. Here comes the idea of ensembling [34]. Ensembling is obtained by combining various models. Bagging, boosting, voting, etc., are some of the ensemble approaches [35]. Here we aggregated predictions of various models, i.e., averaged the output predictions of all models and produced one model closer to the actual effort than any individual model. We took the Linear Regression, Decision Tree Regression, Random Forest Regression, Support Vector Regression, and Neural Network Regression outputs, calculated the average of all the values, and compared them with the actual effort in the test dataset. The results are given for evaluation metrics.

#### F. Pseudo Code

```
# Step 1: Collect historical project data
data = load_data()
# Step 2: Preprocess the data
data = preprocess_data(data)
# Step 3: Train multiple estimation
models
models = []
for i in range(num_models):
    model = train_model(data)
    models.append(model)
# Step 4: Implement the voting algorithm
def ensemble_predict(models, input):
    predictions = [model.predict(input)
for model in models]
    return voting_algorithm(predictions)
# Step 5: Evaluate the ensemble with
voting
validation_set = load_validation_set()
ensemble_accuracy = 0
for input, target in validation_set:
    ensemble_prediction =
ensemble_predict(models, input)
    ensemble_accuracy +=
evaluate_prediction(ensemble_prediction,
target)
ensemble_accuracy /= len(validation_set)
# Step 6: Investigate the impact of
different factors
```

```
# For example, vary the number of models,
the type of models, the voting algorithm,
and the size and quality of the
historical data.
```

```
# Step 7: Apply the ensemble with voting
to new data
new_data = load_new_data()
for input in new_data:
    ensemble_prediction =
ensemble_predict(models, input)
process_prediction(ensemble_prediction)
```

#### IV. EVALUATION CRITERIA

In problems like predicting continuous values, we calculate the error rate given as the difference between the actual and predicted values. For our problem statement, we looked at the MAE (Mean Absolute Error), MSE (Mean Squared Error), and RMSE (Root Mean Square Error) metrics, which are used to compare models.

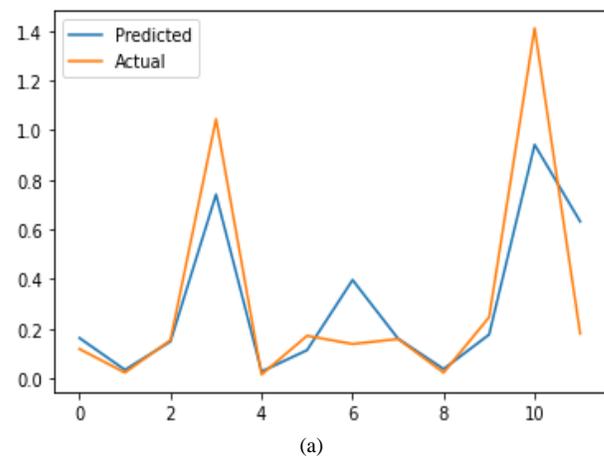
$$MAE = \frac{\sum abs(y_{actual} - y_{predicted})}{n} \quad (2)$$

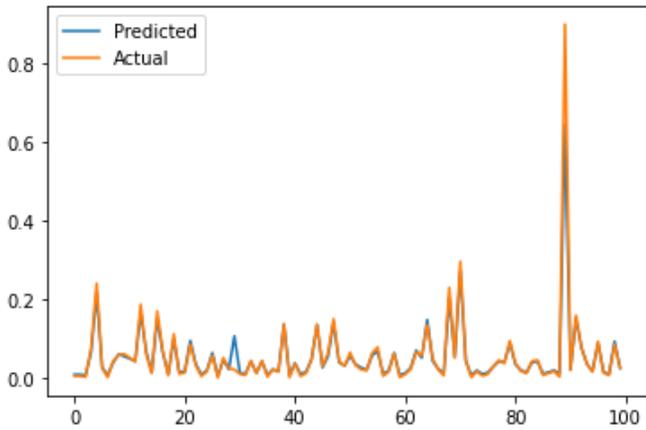
$$MSE = \frac{\sum (y_{actual} - y_{predicted})^2}{n} \quad (3)$$

$$RMSE = \sqrt{\frac{\sum (y_{actual} - y_{predicted})^2}{n}} \quad (4)$$

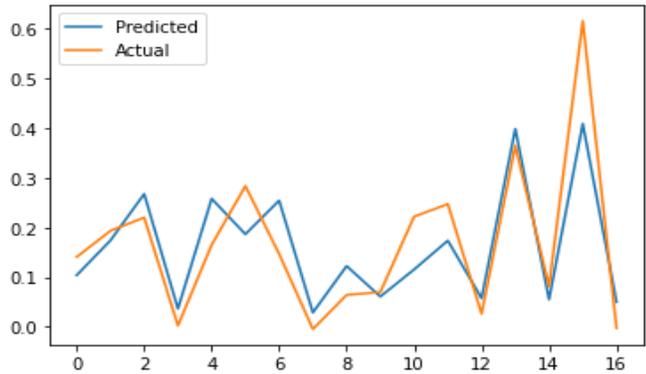
#### V. RESULTS AND DISCUSSION

Below, Fig. 2 represents the deviation between actual and predicted effort values on all datasets, where the X-axis represents the record number. The Y-axis represents the effort of the record. Fig. 2(a), 2(d) on Cocomo81 and Maxwell show a notifiable difference in peak effort values. The values are closer to the China dataset in Fig. 2(b). Fig. 2(c) and 2(f) show a constant gap between actual and predicted values. Fig. 2(e) on Albrecht shows a considerable difference.

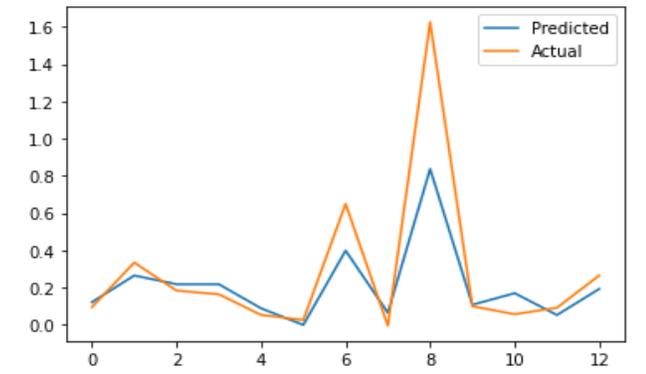




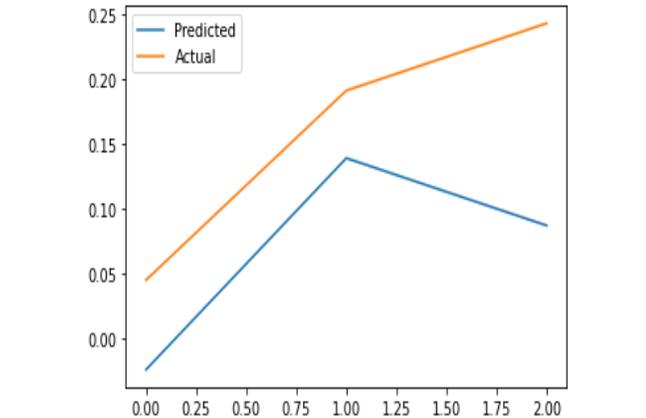
(b)



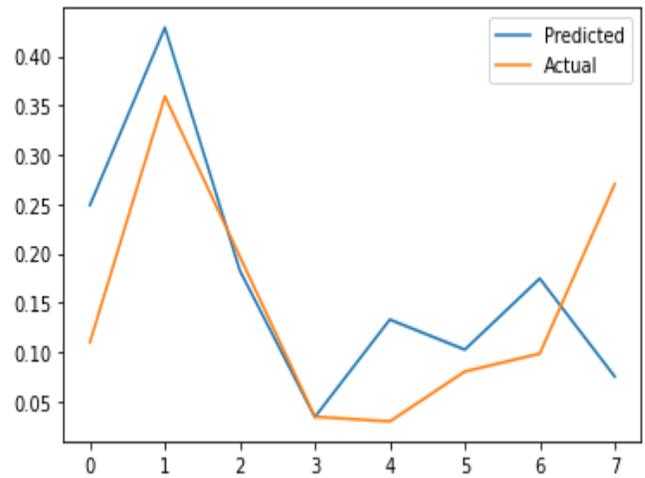
(c)



(d)



(e)



(f)

Fig. 2. (a) Cocomo81 actual vs. predicted effort (b) China's actual vs. predicted effort (c) Desharnais actual vs. predicted effort (d) Maxwell actual vs. predicted effort (e) Kemerer actual vs. predicted effort (f) Albrecht actual vs. predicted effort

Fig. 2(a) represents a line graph drawn to show the deviations between actual effort and the predicted effort by our proposed model in the COCOMO81 dataset. This graph shows a noticeable difference at the peak points.

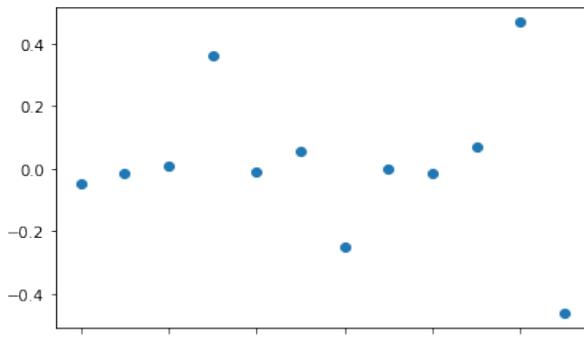
Fig. 2(b) represents a line graph drawn to show the deviations between actual effort and the predicted effort by our proposed model in the China dataset. We can see that the predicted line has come close to the actual line in many places.

Fig. 2(c) represents a line graph drawn to show the deviations between actual effort and the predicted effort by our proposed model in the Desharnais dataset. In this graph, records 10 and 11 have a significant deviation, whereas records 12 to 14 have the least deviation.

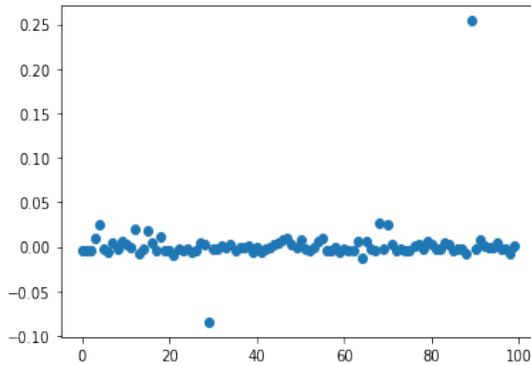
Fig. 2(d) represents a line graph drawn to show the deviations between actual effort and the predicted effort by our proposed model in the Maxwell dataset. This graph shows a noticeable difference at the peak points.

Fig. 2(e) represents a line graph drawn to show the deviations between actual effort and the predicted effort by our proposed model in the Kemerer dataset. As the test set records are meager, they show a significant deviation, but the deviation range is 0.05.

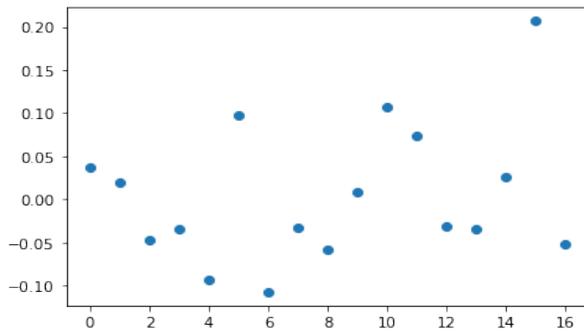
Fig. 2(f) represents a line graph drawn to show the deviations between actual effort and the predicted effort by our proposed model in the Albrecht dataset, where the X-axis represents the record number and the Y-axis represents the effort of the record. Fig. 3 represents the residuals graphs between actual and predicted effort values on all datasets. Fig. 3(a) and 3(d) on Cocomo81 and Maxwell shows a notifiable difference in peak effort values. Fig. 3(b) of the China dataset shows values closer to 0 ("zero"). Fig. 3(c) and 3(f) show a constant gap between actual and predicted values. There is a significant difference in Albrecht's Fig. 3(e).



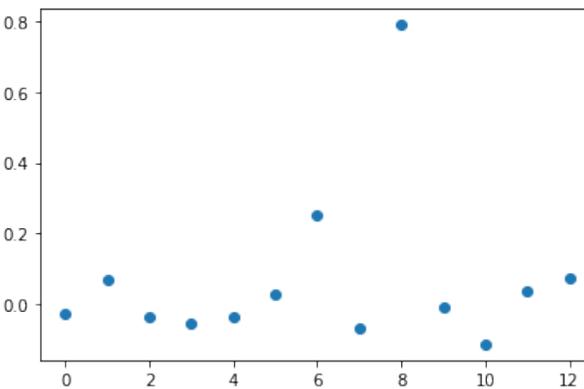
(a)



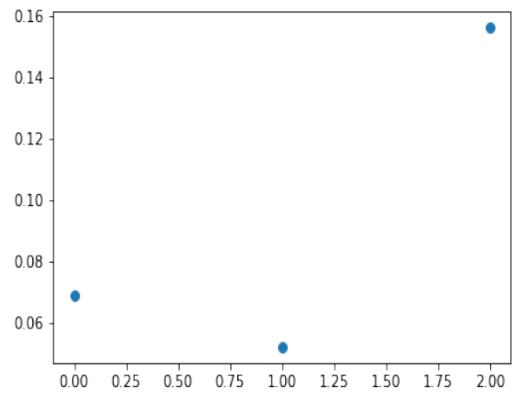
(b)



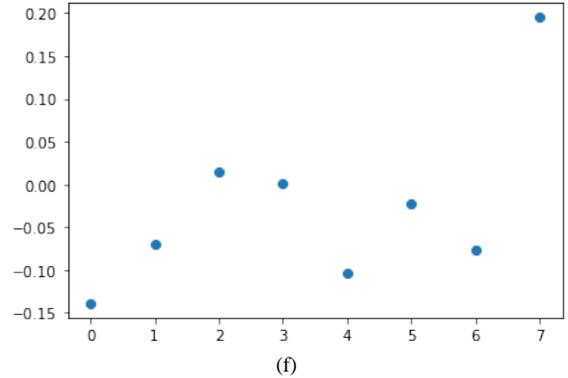
(c)



(d)



(e)



(f)

Fig. 3. (a) Cocomo81 prediction residuals (b) China prediction residuals (c) Desharnais prediction residuals (d) Maxwell prediction residuals (e) Kemerer prediction residuals (f) Albrecht prediction residuals

The above Fig. 3(a) represents a graph that shows the residuals between actual effort and the predicted effort of the data records of the COCOMO81 dataset ranging from -0.4 to +0.4, and most of the data points are present in the range of -0.2 to +0.2.

The above Fig. 3(b) represents a graph that shows the residuals between actual effort and the prediction effort of the data records of the China dataset ranging from -0.10 to +0.25. In the presented graph, most data points are nearer to 0, indicating that the proposed model is working much more efficiently in the China dataset.

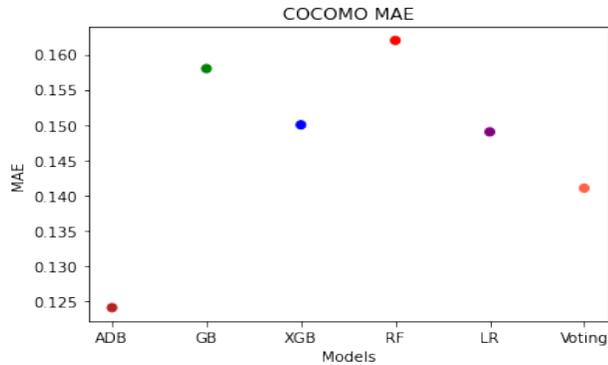
The above Fig. 3(c) represents a graph that shows the residuals between actual effort and the predicted effort of the data records of the Desharnais dataset, ranging from -0.10 to +0.20. In this graph, most of the data points are below point 0. That means the proposed model predicted values are less than the actual values.

Fig. 3(d) shows a graph of the residuals between actual effort and predicted effort of the Maxwell dataset data records, ranging from -0.2 to +0.8. According to this graph, the proposed model prediction is much closer to the actual values based on working on this dataset.

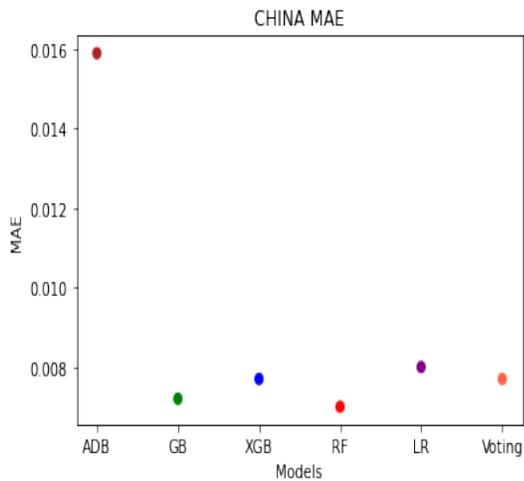
Fig. 3(e) depicts a graph displaying the residuals between actual effort and predicted effort of the Kemerer dataset data records, ranging from 0.05 to +0.16.

The above Fig. 3(f) represents a graph that shows the residuals between actual effort and the predicted effort of the data records of the Albrecht dataset ranging from -0.15 to +0.20.

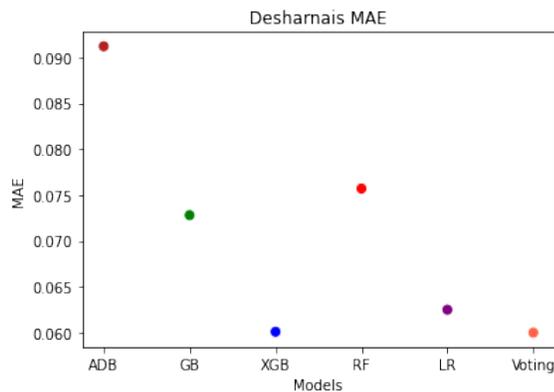
Below, Fig. 4 shows the bar plots of all the implemented models, representing the mean absolute error on all six datasets. Fig. 4(a) the voting model outperformed GB, XGB, RF, and LR except for ADB. Fig. 4(b) shows that, except for RF, voting showed less residual than all others. Fig. 4(c), (d), and (f) voting models are reliable. From all the above comparisons, we concluded that voting is a constant performer. On all datasets, the models behave randomly, whereas voting shows an upvote constantly.



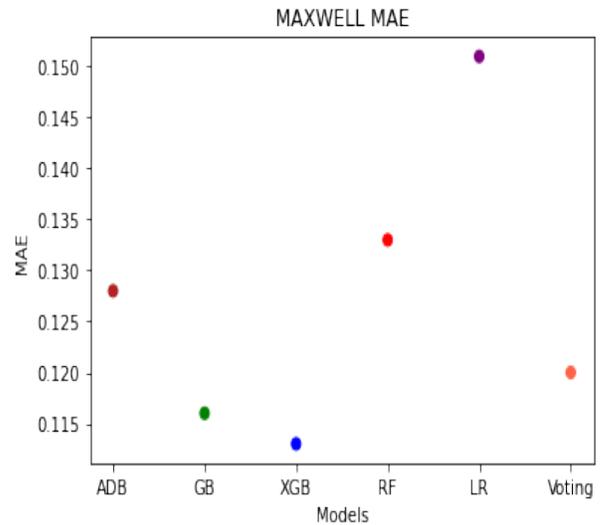
(a)



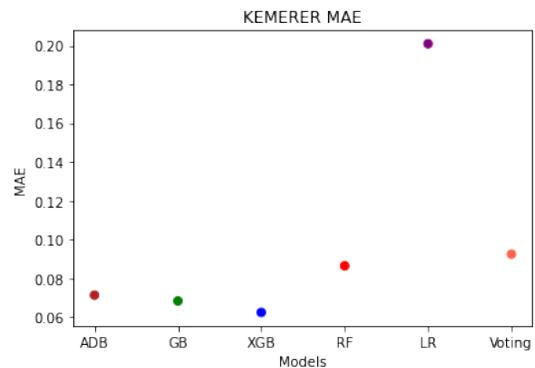
(b)



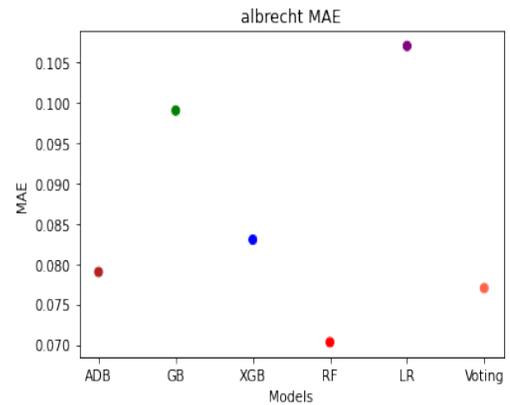
(c)



(d)



(e)



(f)

Fig. 4. (a) COCOMO81 mean absolute error (b) CHINA mean absolute error (c) DESHARNAIS mean absolute error (d) MAXWELL mean absolute error (e) KEMERER mean absolute error (f) ALBRECHT mean absolute error

The graphical representation of Mean Absolute Error for various models that are worked on the COCOMO dataset is shown in Fig. 4(a), with the ADB model giving the slightest error followed by voting and the Random Forest giving the highest error among the models presented.

Fig. 4(b) shows a graphical representation of the Mean Absolute Error for various models tested on the CHINA

dataset. The RF model produces the lowest error, and the ADB produces the highest error.

Fig. 4(c) shows a graphical representation of the Mean Absolute Error for various models tested on the Desharnais dataset, with the Voting and XGB models producing the lowest error and the ADB having the highest error.

The graphical representation of Mean Absolute Error for various models that are worked on the MAXWELL dataset is shown in Fig. 4(d), with the XGB model giving a minor error and the LR model giving the highest error among the models presented.

Fig. 4(e) shows a graphical representation of the Mean Absolute Error for various models tested on the Kemerer dataset. The XGB model produces the lowest error, and the LR model produces the highest error.

Fig. 4(f) shows a graphical representation of the Mean Absolute Error for various models tested on the Albrecht dataset. The RF model produces the lowest error, and the LR model produces the highest error.

We normalized the literature results and compared them with the obtained model's results (see Tables I-VI)

TABLE I. COCOMO81 DATASET

Model	MAE	MSE	RMSE
Linear Regression	0.1499	0.0393	0.1984
AdaBoost	0.1248	0.0395	0.1989
Random Forest	0.1627	0.0680	0.2608
Gradient Boosting	0.1587	0.0662	0.2574
XGB	0.1509	0.0680	0.2665
Ali et al. [10]	0.1652	-	0.4322
Voting	0.1466	0.0527	0.2297

TABLE II. CHINA DATASET

Model	MAE	MSE	RMSE
Linear Regression	0.0080	0.0005	0.0231
AdaBoost	0.0159	0.0013	0.0363
Random Forest	0.0070	0.0008	0.0286
Gradient Boosting	0.0072	0.0008	0.0295
XGB	0.0077	0.0009	0.0308
Hosni et al. [14]	0.0099	-	-
Voting	0.0077	0.0007	0.0270

TABLE III. DESHARNAIS DATASET

Model	MAE	MSE	RMSE
Linear Regression	0.0625	0.0070	0.0841
AdaBoost	0.0912	0.0107	0.1037
Random Forest	0.0757	0.0095	0.0976
Gradient Boosting	0.0728	0.0075	0.0866
XGB	0.0601	0.0065	0.0790
De Carvalho et al., [2]	0.0562	0.0078	0.0880
Hosni et al. [14]	0.0664	-	-
Voting	0.0627	<b>0.0061</b>	<b>0.0783</b>

TABLE IV. MAXWELL DATASET

Model	MAE	MSE	RMSE
Linear Regression	0.1519	0.0571	0.2390
AdaBoost	0.1287	0.0503	0.2243
Random Forest	0.1333	0.0739	0.2719
Gradient Boosting	0.1166	0.0565	0.2378
XGB	0.1131	0.0544	0.2334
Voting	0.1221	0.0555	0.2356

TABLE V. KEMERER DATASET

Model	MAE	MSE	RMSE
Linear Regression	0.2009	0.0462	0.2151
AdaBoost	0.0714	0.0085	0.0922
Random Forest	0.0865	0.0074	0.0865
Gradient Boosting	0.0684	0.0066	0.0815
XGB	0.0625	0.0079	0.0893
Ali et al. [10]	0.1113	-	0.2200
Hosni et al. [14]	0.0866	-	-
Voting	0.0925	0.0160	0.1031

TABLE VI. ALBRECHT DATASET

Model	MAE	MSE	RMSE
Linear Regression	0.1078	0.1977	0.1406
AdaBoost	0.0790	0.0113	0.1064
Random Forest	0.0786	0.0077	0.0878
Gradient Boosting	0.0995	0.0168	0.1298
XGB	0.0835	0.0115	0.1073
Ali et al. [10]	0.0856	-	0.1196
Voting	<b>0.0775</b>	0.0099	0.0996

Our work includes testing the voting regressor on six datasets. From the above tables, observations in all datasets voted on, showed excellent performance in minimizing the actual and predicted effort error. On the COMO81 dataset, absolute error is the minimum for voting, and squared errors are minor for linear regression. On COCOMO81, China, Desharnais, Kemerer, Maxwell, and Albrecht had excellent performances. Finally, we concluded that all dataset implementations support voting, which makes voting more reliable and robust. Voting followed by linear regression shows that the datasets have a linear relationship between the attributes of the projects.

## VI. CONCLUSION

We studied various existing research papers on software effort estimation in this work. In the early days, we relied on many conventional approaches, considering the line of codes, functional points, CPM and PERT, etc., or merely relying on the people's judgment that has ample experience in software project effort determination. Because extensive developments in project building consider multiple parameters in every project, these techniques might not be feasible anymore with rapid results in software projects. And at the same time, machine learning has gained momentum in recent decades in various domains. And there is some work taking place in software engineering through machine learning. Therefore, our work aims to provide a robust machine learning model for effort calculation. We successfully used the machine learning ensembling concept to predict software development efforts. We considered every parameter for the effort estimation. Based on our research, the ensembling of models outperformed other single models. We recorded a lower error rate from the ensemble model comparatively. The average of different predictors positively impacted the output, which shows the vital role played in optimizing software effort estimation in the machine learning area. The input dataset dramatically affects how well the machine learning algorithm works, and in our work, models performed very well with our datasets.

## REFERENCES

- [1] Ramesh, M. R., & Reddy, C. S. (2016). Difficulties in software cost estimation: A survey. *International Journal of Scientific Engineering and Technology*, 5(1), 10-13.
- [2] Hareton, L., and Zhang F: "Software Cost Estimation", Department of Computing, Hong Kong Polytechnic University, <http://paginaspersonales.deusto.es/cortazar/doctorado/articulo/s/leung-andbook.pdf>, accessed 24th Nov 2019.
- [3] Rajeswari, K., & Beena, D. R. (2018). A Critique on Software Cost Estimation. *International Journal of Pure and Applied Mathematics*, 118(20), 3851-3862.
- [4] Bull Survey, 1998, Failure Causes. [http://www.itcortex.com/Stat\\_Failure\\_Cause.htm#surveys](http://www.itcortex.com/Stat_Failure_Cause.htm#surveys), Retrieved on 1st June 2013.
- [5] KPMG Canada, 1997, Failure Causes. [http://itcortex.com/Stat\\_Failure\\_Cause.htm#surveys](http://itcortex.com/Stat_Failure_Cause.htm#surveys). Retrieved on 2nd June 2013.
- [6] Ziauddin, Shahid Kamal Tipu, ShahrughZia, "An Effort Estimation Model for Agile Software Development", *Advances in Computer Science and its Applications (ACSA)*, Vol. 2, No. 1, 2012, ISSN 2166-2924.
- [7] Boehm, B. W. (2017, May). Software cost estimation meets software diversity. In 2017 IEEE/ACM 39th International Conference on Software Engineering Companion (ICSE-C) (pp. 495-496). IEEE.
- [8] Santanu Kumar Rath, "Use Case Point Approach Based Software Effort Estimation using Various Support Vector Regression Kernel Methods", January 2014.
- [9] Ali BouNassif, Mohammad Azzeh, Ali Idri, and Alain Abran, *Hindawi, Software Development Effort Estimation Using Regression Fuzzy Models, Computational Intelligence and Neuroscience Volume 2019*.
- [10] C. E. L. Peixoto, J. L. N. Audy and R. Prikladnicki, "Effort Estimation in Global Software Development Projects: Preliminary Results from a Survey," 2010 5th IEEE International Conference on Global Software Engineering, Princeton, NJ, 2010, pp. 123-127, doi: 10.1109/ICGSE.2010.22.
- [11] B Rajesh Kumar Singh, A.K.Misra, "Software Effort Estimation by Genetic Algorithm Tuned Parameters of Modified Constructive Cost Model for NASA Software Projects", *International Journal of Computer Applications* 59(9):22-26, December 2012.
- [12] Kumar, G., & Bhatia, P. K.. Empirical assessment and optimization of software cost estimation using soft computing techniques. In *Advanced Computing and Communication Technologies* (pp. 117-130). Springer, Singapore, 2016.
- [13] Ahmad, S. W., & Bamnote, G. R.. Whale-crow optimization (WCO)-based Optimal Regression model for Software Cost Estimation. *Sādhanā*, 44(4), 94, 2019.
- [14] Amini, F., & Hu, G. (2021). A two-layer feature selection method using a genetic algorithm and elastic net. *Expert Systems with Applications*, 166, 114072.
- [15] De Carvalho, H. D. P., Fagundes, R., & Santos, W. (2021). Extreme Learning Machine Applied to Software Development Effort Estimation. *IEEE Access*, 9, 92676-92687.
- [16] Ghosh, K. K., Ahmed, S., Singh, P. K., Geem, Z. W., & Sarkar, R. (2020). Improved binary sailfish optimizer based on adaptive  $\beta$ -hill climbing for feature selection. *IEEE Access*, 8, 83548-83560.
- [17] Carbonera, C. E., Farias, K., & Bischoff, V. (2020). Software development effort estimation: a systematic mapping study. *IET Software*, 14(4), 328-344.
- [18] Chhabra, S., & Singh, H. (2020). Optimizing design of a fuzzy model for software cost estimation using particle swarm optimization algorithm. *International Journal of Computational Intelligence and Applications*, 19(01), 2050005.
- [19] Ghatasheh, N., Faris, H., Aljarah, I., & Al-Sayyed, R. M. (2019). Optimizing software effort estimation models using the firefly algorithm. *arXiv preprint arXiv:1903.02079*.
- [20] Wani, Z. H., & Quadri, S. M. K. (2019). An improved particle swarm optimization-based functional link artificial neural network model for software cost estimation. *International Journal of Swarm Intelligence*, 4(1), 38-54.
- [21] Ali, A., & Gravino, C. (2019, December). Using Combinations of Bio-inspired Feature Selection Algorithms in Software Efforts Estimation: An Empirical Study. In 2019 13th International Conference on Open Source Systems and Technologies (ICOSST) (pp. 1-8). IEEE.
- [22] Kodmelwar, M. K., Joshi, S. D., & Khanna, V. (2018). A deep learning modified neural network is used for efficient effort estimation. *Journal of Computational and Theoretical Nanoscience*, 15(11-12), 3492-3500.
- [23] Desai, V. S., & Mohanty, R. (2018, October). ANN-Cuckoo optimization technique to predict software cost estimation. In 2018 Conference on Information and Communication Technology (ICT) (pp. 1-6). IEEE.
- [24] Langsari, K., & Sarno, R. (2018). Optimizing effort parameter of COCOMO II using particle swarm optimization method. *Telkomnika*, 16(5), 2208-2216.
- [25] Hosni, M., Idri, A., Abran, A., & Nassif, A. B. (2018). On the value of parameter tuning in heterogeneous ensembles effort estimation. *Soft Computing*, 22(18), 5977-6010.
- [26] Goyal, S., & Parashar, A. (2018). Machine learning application to improve COCOMO model using neural networks. *International Journal of Information Technology and Computer Science (IJITCS)*, 3, 35-51.

- [27] Padhy, N., Singh, R. P., & Satapathy, S. C. (2018). Software reusability metrics estimation: algorithms, models and optimization techniques. *Computers & Electrical Engineering*, 69, 653-668.
- [28] Pospieszny, P., Czarnacka-Chrobot, B., & Kobylinski, A. (2018). An effective approach for software project effort and duration estimation with machine learning algorithms. *Journal of Systems and Software*, 137, 184-196.
- [29] Shekhar, S., & Kumar, U. (2016). Review of various software cost estimation techniques. *International Journal of Computer Applications*, 141(11), 31-34.
- [30] Venkataiah, V., Mohanty, R., Pahariya, J. S., & Nagaratna, M. (2017). Application of ant colony optimization techniques to predict software cost estimation. In *Computer Communication, Networking and Internet Security* (pp. 315-325). Springer, Singapore.
- [31] Nassif, A. B., Azzeh, M., Capretz, L. F., & Ho, D. (2016). Neural network models for software development effort estimation: a comparative study. *Neural Computing and Applications*, 27(8), 2369-2381.
- [32] Miandoab, E. E., & Gharehchopogh, F. S. (2016). A novel hybrid algorithm for software cost estimation based on cuckoo optimization and k-nearest neighbors algorithms. *Engineering, Technology & Applied Science Research*, 6(3), 1018-1022.
- [33] Dizaji, Z. A., & Gharehchopogh, F. S. (2015). A hybrid of ant colony optimization and chaos optimization algorithms approach for software cost estimation. *Indian Journal of science and technology*, 8(2), 128.
- [34] Mahmood, Y., Kama, N., Azmi, A., Khan, A. S., & Ali, M. (2021). Software Effort Estimation Accuracy Prediction of Machine Learning Techniques: A Systematic Performance Evaluation. *ArXiv*. <https://doi.org/10.48550/arXiv.2101.10658>
- [35] Marco, R., Ahmad, S. S. S., & Ahmad, S. (2022). Bayesian hyperparameter optimization and Ensemble Learning for Machine Learning Models on software effort estimation. *International Journal of Advanced Computer Science and Applications*, 13(3).

# An Effective Heart Disease Prediction Framework based on Ensemble Techniques in Machine Learning

Deepali Yewale<sup>1</sup>, S. P. Vijayaragavan<sup>2</sup>, V. K. Bairagi<sup>3</sup>

Department of Electronics and Communication, Bharath Institute of Higher Education and Research, Chennai-611026, India<sup>1</sup>

Department of Electrical and Electronics Engineering, Bharath Institute of Higher Education and Research, Chennai-611026, India<sup>2</sup>

Department of Electronics and Telecommunication, AISSMS Institute of Information Technology, Pune-411001, India<sup>1,3</sup>

**Abstract**—To design a framework for effective prediction of heart disease based on ensemble techniques, without the need of feature selection, incorporating data balancing, outlier detection and removal techniques, with results that are still at par with cutting-edge research. In this study, the Cleveland dataset, which has 303 occurrences, is used from the UCI repository. The dataset comprises 76 raw attributes, however only 14 of them are listed by the UCI repository as significant risk factors for heart disease when the dataset is uploaded as an open source dataset. Data balancing strategies, such as random over sampling, are used to address the issue of unbalanced data. Additionally, an isolation forest is used to find outliers in multivariate data, which has not been explored in previous research. After eliminating anomalies from the data, ensemble techniques such as bagging, boosting, voting, stacking are employed to create the prediction model. The potential of the proposed model is assessed for accuracy, sensitivity, and specificity, positive prediction value (PPV), negative prediction value (NPV), F1 score, ROC-AUC and model training time. For the Cleveland dataset, the performance of the suggested methodology is superior, with 98.73% accuracy, 98% sensitivity, 100% specificity, 100% PPV, 97% NPV, 1 as F score, and AUC as 1 with comparatively very less training time. The results of this study demonstrate that our proposed approach significantly outperforms the existing scholarly work in terms of accuracy and all the stated performance metrics. No earlier research has focused on these many performance parameters.

**Keywords**—Machine learning; heart disease; ensemble techniques; random over sampling, isolation forest

## I. INTRODUCTION

Cardiovascular disease (CVD) is considered to be the foremost reason of death in the world. It is estimated that half of all CVD cases occur in Asia. Besides, near about three-quarters of all the global mortality are anticipated to happen because of persistent diseases by 2021, with 75% of deaths due to heart disease. These circumstances create an imperative need to design decision support system (DSS) for early prediction of heart disease. About 80 % of CVD are preventable if predicted at an early stage.

Traditional method of heart disease diagnosis includes extensive examination of patient. In this case the diagnosis of disease totally depends upon domain experts' knowledge and exactness of collected clinical data. This invasive method of diagnosis is not reliable and efficient. So there is need for cost effective and highly accurate Non-invasive approach such as Machine Learning (ML) to design prediction model.

ML is able to filter out pertinent connection between tremendous measures of data. Advancement in Artificial intelligence has played very significant role in the new era of computing. Various researchers have tried and tested data mining techniques in the healthcare domain and found it to be outperforming. The limitation of the prevailing analysis was the utilization of tedious task of feature engineering within the classification process. The principal inspiration of this research is to makeover data to usable form so that researchers can use it to design DSS to enhance safety of the patients.

In this work, proposed a model with data pre-processing, data balancing and outlier detecting followed by ensemble classifier. Standard scalar is used in the data pre-processing stage where each attributes with standard deviation of one is achieved. Imbalanced data is explored and oversampling technique was proposed to improve the reliability of the model. Anomalies or outliers are considered as noise in the data and may lead to misclassification. Furthermore proposed anomalies detection and removal using Isolation forest algorithm. Besides, Ensemble techniques bagging, boosting, Voting and stacking are applied to measure the effectiveness of the proposed methodology. This article focuses on the prediction of cardiac disease employing ensemble techniques of ML without feature engineering.

The main contributions of the proposed work include:

- A novel combination of data standardization, data balancing, and outlier detection to transform the data into usable form.
- The study involves the isolation forest for outlier detection of multivariate data, which has not been extensively explored in the previous research.
- The study assesses the performance of heart disease prediction system using ensemble techniques without feature selection designed, which have not been studied in depth in previous research.

The rest of the paper is structured as follows: In Section II literature analysis is provided, research gap is highlighted in Section III; Methodology is projected in Section IV. The detailed results of the proposed approach are presented in Section V followed by comparative study with existing research work in Section VI, and finally conclusion and future scope in Section VII.

## II. LITERATURE ANALYSIS

According to WHO, heart disease represent predominant reason of death in developing countries. One of the reasons to fail in the treatment of heart disease is unidentified pattern with cardiac data. Machine learning has been proved the remedy for that, as it is able to extract the pattern in cardiac data to predict the heart disease.

Researchers have explored and evaluated several methodologies, including single-base classifiers, ensemble approaches, and hybrid techniques as the prediction model. Furthermore, data pre-processing techniques, feature selection techniques, and optimization approaches were employed to improve the performance of the prediction system.

Many researchers implemented basic ML classifiers on a cardiac dataset and achieved good results. Authors suggested a modified random forest [1] to boost the prediction ability of the classifier. The proposed work achieved the highest accuracy of 86.84% with the UCI Cleveland dataset. Author [2] implemented Logistic regression (LR), K nearest neighbour (KNN) and Random Forest (RF) classifiers on a medical dataset from the UCI repository. The highest accuracy achieved with KNN is 87.5%, when implemented on the Python platform.

The above mentioned, state of the art research used conventional algorithms to design decision support system for heart disease prediction and it has been observed that the average accuracy is below 90%.

Instead of relying on a conventional model to provide an exceptional solution, the ensemble method leverages the strengths of numerous models to mitigate the limitations of a single model. In [3], proposed homogeneous ensemble learning using an accuracy-based weighted ageing classifier. The proposed model achieved an accuracy of 93% on the Cleveland dataset. Instead of utilizing conventional single model, majority vote Ensemble model can be used [4] in the prediction system of heart disease. This approach has produced 90% accuracy for Cleveland dataset. According to this literature survey, an ensemble method has shown to be more successful than a single model strategy.

Many feature selection methods have been proposed by researchers for obtaining more relevant features from a given dataset. Javeed et al. [5] proposed a randomized search algorithm (RSA) to get the optimal subset of features, and grid search optimized RF was used as a classifier. The experimental results have achieved 93.33% accuracy while improving the training accuracy as well. Muhammad et al. [6] proposed a model where four feature selection methods, namely fast correlation-based feature selection (FCBF), minimum redundancy, maximal relevance (mRMR), least absolute and selection operator (LASSO), and Relief, were tested on 10 different machine learning classifiers. It was found that, for the features chosen by FCBF, ETC's accuracy increased from 92.09 % to 94.41 % compared to complete features. Dissanayake et al. [7] performed research where filter, wrapper, and embedded feature selection approaches were

implemented. The test findings show that DT with backward elimination wrapper feature selection outperforms with an accuracy of 88.52 %.

## III. RESEARCH GAP

The majority of the scholarly work is focused on improving accuracy via feature selection techniques. However, to eliminate the data cleaning operations while yielding high disease prediction accuracy, a computationally effective feature selection approach is required [8]. There is need to investigate a new intelligent technique to generate a meaningful concise set of features. Noise and outliers present in the data make it difficult to select exact features [9]. As feature selection is a tiresome activity and only some of the existing work discussed in the literature is able to predict heart disease with good accuracy, there is a dire need to test machine learning framework without feature selection for the effective prediction of cardiovascular disease [10].

The dataset discussed in the existing work is imbalanced with an uneven contribution of the majority and minority classes. Class imbalance has not been amply focused in the previous research. The problem of class imbalance must be taken care of before implementing any classification mechanism. On the other hand; only a few researchers have worked on outlier removal from dataset. There is a need for an outlier detection method that can differentiate between normal data and outliers [11]. Only few Researchers have experimented unsupervised outlier detection techniques such as DBSCAN, isolation forest, K-means clustering. There is need to experiment and analyse Isolation forest for outlier detection. The utilization of ensemble based algorithms needs to be experimented rigorously and analysed for effective prediction of heart disease.

The unique aspect of the proposed research work is to design a framework for heart disease prediction that can handle the problems of imbalanced class, outlier detection and still conveys comparable performance index without any feature engineering.

## IV. MATERIALS AND METHODS

In this research, we present a new paradigm for predicting cardiac disease, which can predict the presence and absence of heart disease reliably, as shown in Fig. 1.

### A. Dataset Description

The Cleveland dataset used for this proposed work has 303 instances with 76 clinical and physical parameters. Most of the research work has chosen just 14 features in their scholarly work as these attributes are the most significant in the prediction of heart disease. Other attributes such as exercise protocol and time when ST measure depression was performed, had minor effects on heart disease and so 62 attributes are omitted by researchers.

As seen in Table I, the UCI repository specifically mentions these 14 attributes when uploading the dataset for open access.

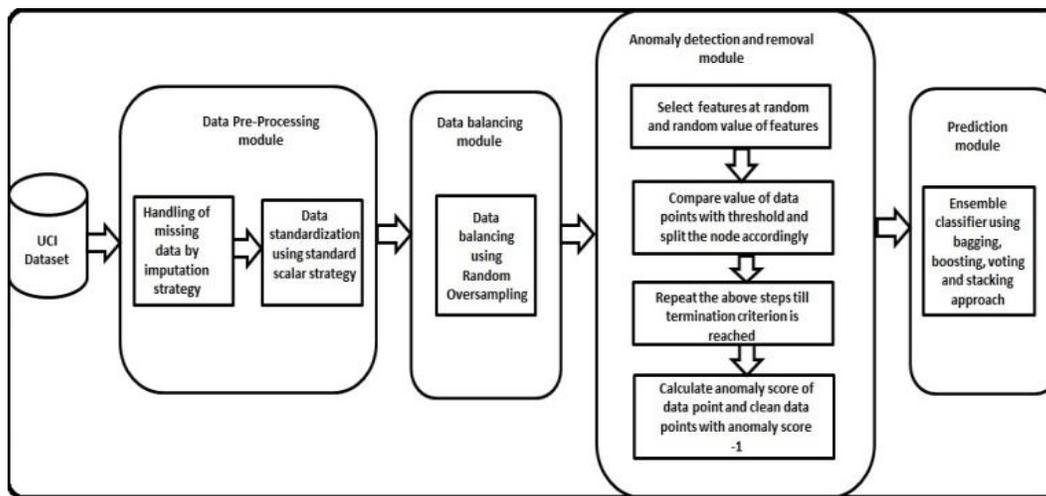


Fig. 1. Proposed heart disease prediction system using random oversampling, isolation forest and ensemble prediction model.

TABLE I. IMPORTANT 14 ATTRIBUTES FROM 76 ATTRIBUTES OF UCI DATASET

Sr. No.	Name of Attribute	Position in the dataset	% of the data Complete
1	Age in years	3	100
2	Sex	4	100
3	Chest pain type	9	100
4	Resting Blood Pressure	10	100
5	Serum Cholesterol	12	100
6	Fasting blood sugar	16	100
7	Resting ECG	19	100
8	Maximum Heart Rate	32	100
9	Exercise-induced angina	38	100
10	ST depression	40	100
11	The slope of the ST segment	41	100
12	Number of containers colored by fluoroscopy	44	98.67
13	Thalassemia	51	99.33
14	Diagnosis value	58	100

Among these, 13 are independent variables and 1 is a dependent target variable for the diagnosis of heart disease, where 0 represents the absence of heart disease and 1 represents the presence of heart disease.

### B. Data Preprocessing

Before catering data into the machine learning classifier, it is important to analyse and pre-process the data to improve its quality. A few attributes, as shown in Table I, have missing values. Missing values are replaced by the mean value of those attributes [12].

Creating a data-analysis-based decision support system necessitates standard data, which frequently necessitates pre-processing activities such as data cleansing, pruning, and scaling. Data standardisation is performed to scale each feature

to unit variance. Attributes assessed at different scales do not contribute equally to the model fitting and may result in bias. To address this possible issue, feature-wise standardisation is utilised prior to model fitting [13]. The feature in each column of  $x$  is normalized independently, so that each feature has a mean  $\mu = 0$  and a standard deviation  $\sigma = 1$ .

A value is standardized as (1).

$$z = \frac{x - \mu}{\sigma} \quad (1)$$

Where mean  $\mu$  is defined in (2).

$$\mu = \frac{1}{N} \sum_{i=1}^N X_i \quad (2)$$

And standard deviation  $\sigma$  is as in (3).

$$\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2} \quad (3)$$

Where  $N$  is the number of instances in each column ( $N=303$ ).

### C. Data Balancing

As per the exploratory data analysis of the target variable, the given Cleveland dataset has 138 instances of healthy people and 165 instances of people with heart disease. The dataset has unequal distribution of negative (0) and positive (1) instances.

Because of this unequal number of positive and negative classes, it will be difficult for machine learning models to learn the pattern of the dataset and it may hamper the performance of the model [14]. To address the problem of imbalanced data, the random oversampling technique is proposed. It randomly duplicates examples from the minority class with substitution and adds them into the training dataset. The superiority of random oversampling is that all individuals from the minority and majority classes are maintained; therefore, no information from the original training set is lost. Synthetic Minority Oversampling Technique (SMOTE) is another method of data for oversampling where samples are created synthetically but it may create noise with high dimensional data. Data augmentation using SMOTE may provide diverse results and may not always be beneficial for medical data [15].

### D. Anomaly Detection

Anomaly detection is the technique of identifying outliers in data. Researchers have preferred to use unsupervised anomaly detection models. Isolation forest is a unique method based on this isolation property of outliers and is fundamentally different from other density-based and cluster-based outlier detection methods [16]. In this research paper, it is proposed to use an isolation forest (iForest) to isolate the anomalies from the data samples.

Here is the algorithm to compute an isolation tree:

- 1) Choose a feature at random from data and refer it as  $f$ .
- 2) Choose a value at random from the feature  $f$  and utilize as threshold ' $t$ '.
- 3) Data points with  $f < t$  are saved in Node 1 whereas the data points with  $f \geq t$  kept in Node 2.
- 4) Steps 1–3 repeated for Node 1 and Node 2.
- 5) Stop the process when the tree has reached full maturity or when a termination requirement is fulfilled.

An isolation tree can be extended to an isolation forest—an ensemble of multiple isolation trees.

The isolation forest in sklearn has 2 important inputs:

$n\_estimators$ : Number of Isolation trees to be trained

$Contamination$ : Fraction of anomalous data points.

In our case we suspect 5% of the data to be anomalous and set contamination to 0.05.

Steps in building an Isolation forests:

- 1) Construct an Isolation Tree either from the entire feature set or a randomly chosen subset of the feature set.
- 2) Construct  $n$  such Isolation trees.
- 3) Calculate an Anomaly score for each data point using formula in (4).

$$s(x, n) = 2^{-E(\hat{h}(x)/c(n))} \quad (4)$$

$s$  = score (closer to 1: outlier, closer to 0: normal data point),

$E(\hat{h}(x))$  = Average path length taken by data point  $x$ ,

$c(n)$  = Average path length of every terminal nodes.

Isolation forest can be used for univariate as well as multivariate dataset. Let us consider our case of the multivariate dataset as shown in Fig. 2.

Isolation tree divides the data into “boxes”. It has the property that it segregates the region containing anomalies earlier than the boxes containing normal data points. If the feature has an anomaly, the anomalous point will be far away from the normal points in the data. It helps isolation forests to isolate out anomalies relatively early in the splitting process.

As shown in the figure, the anomaly can be detected at split 2 only. If we go on splitting the data, few normal points got isolated much later as shown in split 4. Isolation Forest can detect the outliers faster and require less memory as compared to other algorithms.

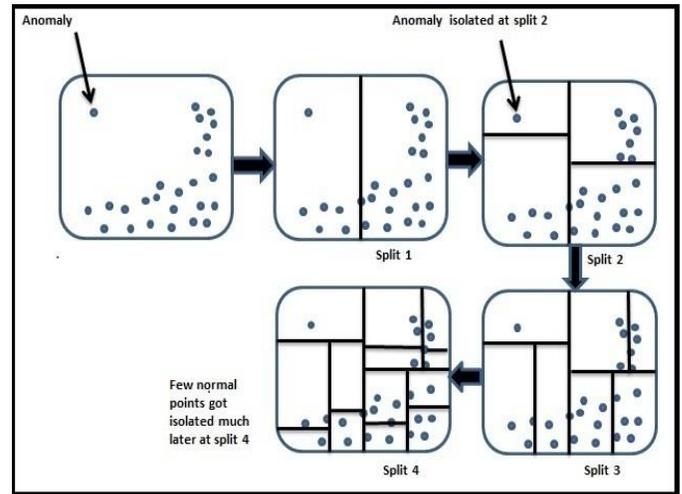


Fig. 2. Splitting process in isolation forest where anomalies are isolated early in split 2 as compared to normal data point isolated in split 4.

### E. Prediction Module: Ensemble ML Techniques

Ensemble learning is a machine learning model in which numerous models (commonly referred to as weak learners or base models) are trained to handle the same issue and then integrated to provide improved results. We tested our model using bagging (RF, ETC), boosting (XGBoost, AdaBoost), voting (hard, soft voting), and staking (RF+SVM) ensemble techniques.

Here we discussed the ensemble techniques, implemented in our proposed work.

**Bagging:** In this kind of approach, several instances of the same base model are trained in parallel (independently from each other) on separate bootstrap samples and then aggregated in some form of "averaging" process.

We have implemented Random Forest and Extra Tree classifier as bagging techniques.

**Random Forest:** This classifier belongs to the ensemble classifier family. It employs decision tree models to improve prediction outcomes. It generates numerous trees from the training data set, and a bootstrap approach is used to each tree. The random forest technique is a bagging method in which deep trees fitted on bootstrap samples are blended to create an output with lower variance.

**Extra Tree classifier:** Extra Trees classifier is a type of ensemble technique that delivers classification result by accumulating the results of multiple uncorrelated decision trees grouped together in a "forest." It differs as compared to Random forest in a way the decision tree is built in the forest. It separates nodes by selecting cut-points completely at random, and it also grows the trees using the whole training sample.

**Boosting:** Boosting is an ensemble modeling strategy that seeks to construct a strong classifier from a collection of weak classifiers to reduce training errors. A random sample of data is chosen, fitted with a model, and then trained sequentially—that is, each model attempts to compensate for the shortcomings of its predecessor. In this work, proposed to use XGBoost and AdaBoost boosting techniques.

**XGBoost:** XGBoost is the state of the art gradient boosted tree algorithm that boosts the performance of weak learners. It uses greedy algorithm to calculate the best split. To begin, a weak classifier is fitted to the data. It adds another weak classifier to upgrade the present model's performance, without losing the prior classifier's performance. The same process is continued and it employs a gradient descent approach to reduce the loss when adding new models. Each new classifier must take into account where the prior classifiers failed to perform well. To generate a new model, the method constantly reduces the errors of prior models in the gradient direction.

**AdaBoost:** AdaBoost is a boosting machine learning algorithm that use weighted linear combination to cascade numerous weak learners into a particular classifier. AdaBoost uses a learning technique to re-weight samples of the original training data in a sequential manner. It is an iterative approach, with each iteration giving more weight to the misclassified occurrences from the preceding iteration. Each instance is initially allocated an identical weight, and iteratively, the weights of all wrongly classified instances are increased while the weights of successfully classified examples are decreased. The algorithm recursively applies the base classifier with fresh weights to the training data. The final classification model produced is a linear combination of all the models developed over the rounds. AdaBoost completely considers each classifier's weight; nonetheless, it is vulnerable to outliers and noisy data.

**Voting Classifier:** Voting classifier aggregates the output of each classifier provided to it and produces the final prediction of the class label of a new instance based on voting. The voting can be of two types, hard or soft. Simple majority voting is utilized in the situation of hard voting. In this situation, the class with the highest number of votes is projected. A forecast is created for soft voting by averaging the class-probabilities of each classifier. The projected class is the one with the highest average probability. In the proposed work, model is checked for both hard and soft voting. In the proposed model LR, NB, DT, SVM, KNN has been ensemble as base models in hard voting and LR, NB, DT, KNN are used in soft voting.

**Stacking:** The stacking approach is a two-layered ensemble technique. The top layer comprises of all the baseline models used to predict the outcomes on the test dataset. The second layer consists of a Meta-Classifier, which accepts all of the baseline model outcomes as input and generates new prediction. The second layer combines the output of the first layer. Here, RF is used as baseline model and SVM as Meta Classifier.

## V. RESULTS AND ANALYSIS

The experiment has been conducted on Python platform using different libraries on an Intel Core i5 processor 9300H CPU with 2.40GHz, 4GB NVIDIA GTX 1650 graphical processing unit Lenovo machine equipped with 8GB RAM. Exploratory data analysis and data pre-processing have been performed. The dataset is divided into 75% of training data and 25% of testing data. Stratified k fold has been introduced in

the dataset's training phase to avoid sampling bias. The accuracy, precision, recall, sensitivity, specificity, PPV, NPV, F1 score, ROC\_AUC score, and computing time of the model are used to validate the performance of the suggested technique.

The experiment focused on the evaluation of the model by implementing random oversampling and isolation forest for various ensemble classifiers. The number of instances after implementing Random oversampling to Cleveland dataset is as shown in Table II.

Before implementation of random oversampling the total number of instances in the dataset are 303 with 138 instances indicating absence of heart disease and 165 instances for presence of heart disease. After processing data for random oversampling, the total number of samples is 330 since the positive and negative class instances are evenly distributed and equal to 165.

TABLE II. UCI CLEVELAND DATASET OVER SAMPLING RESULTS

Class	Absence of HD	Presence of HD
Before Random Oversampling	138	165
After Random Oversampling	165	165

The isolation forest is used to identify and clear outliers from the dataset. After removing outliers, the dataset has 313 samples. The performance assessment of the ensemble classifier model using the proposed methodology is shown in Table III.

It demonstrates that the implementation of random oversampling and an isolation forest has given excellent performance for all the ensemble classifiers. Many existing research papers put emphasis on feature engineering, but our proposed approach works on the whole featured dataset and demonstrates the significance of data balancing and outlier removal in the prediction of heart disease. The suggested approach outperformed numerous existing studies in the literature without requiring high computational time. The accuracy of the proposed model for ensemble classifiers is in the range of 97.47 % to 98.73 % for the Cleveland dataset. The highlighted column demonstrates that soft voting provides excellent result among all the implemented ensemble methods.

Confusion matrix for implemented ensemble classifiers is as shown in Fig. 3.

From the confusion matrix, it has been observed that, RF, ETC and AdaBoost Ensemble Techniques provide False Negative (FN) value of 2, that means two patients with actual heart disease are incorrectly predicted as non-heart disease persons. For XGBoost, Voting and hybrid ensemble techniques, FN value is 1 indicating only one heart disease patient incorrectly predicted as non-heart disease patient by the model.

Fig. 4 explores the graphical presentation of all the performance parameters of the proposed methodology with all the ensemble techniques implemented.

TABLE III. PERFORMANCE EVALUATION OF PROPOSED METHODOLOGY FOR CLEVELAND DATASET

Metric	RF	ETC	XGBoost	AdaBoost	Hard Voting	Soft Voting	Stacking
Accuracy	97.47	97.47	98.73	97.47	98.73	98.73	98.73
Sensitivity	95	95	98	95	98	98	98
Specificity	100	100	100	100	100	100	100
PPV	100	100	100	100	100	100	100
NPV	95	95	97	95	97	97	97
F1-Score	0.97	0.97	0.99	0.97	0.99	0.99	0.99
AUC	0.998	0.982	0.995	0.995	--	1	0.980
Computational Time in sec.	0.054	0.062	0.144	0.114	0.016	0.011	0.631

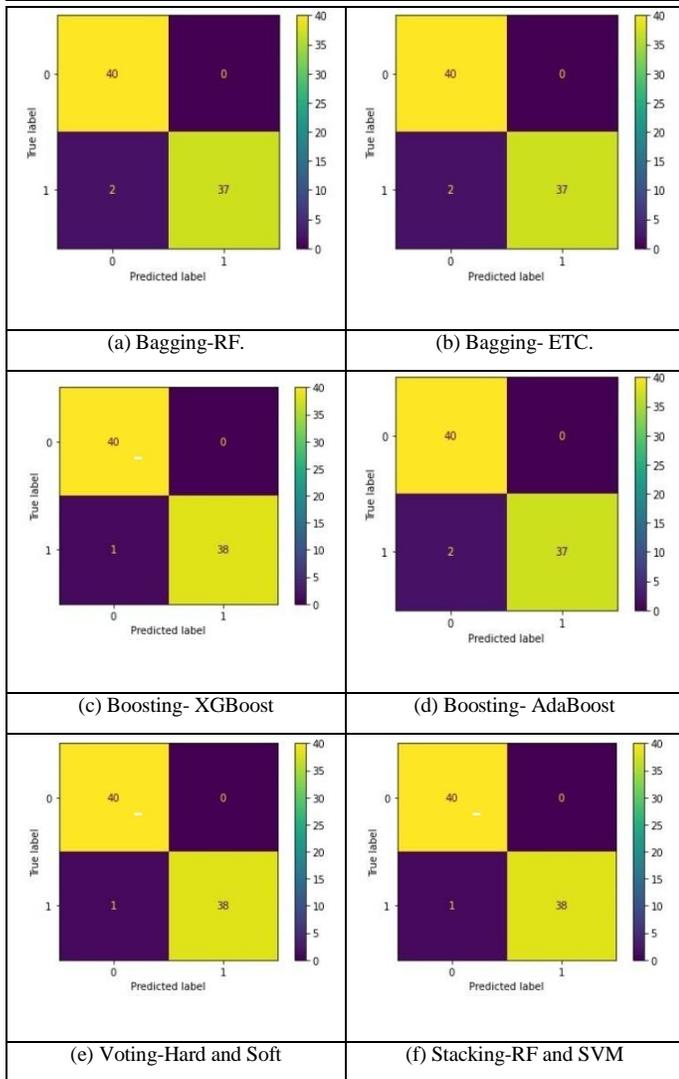


Fig. 3. Confusion matrix of implemented ensemble classifiers.

It reveals that all the ensemble techniques provide excellent performance when applied to random over sampling to cater with imbalance data and isolation forest for anomaly detection to the Cleveland dataset with 303 instances without any feature selection.

Proposed work is additionally assessed for ROC\_AUC analysis. The ROC curve is a depiction of the True positive

rate vs the False positive rate. In other words, it is trade-off between sensitivity and specificity. The area the ROC curve (AUC) is said to be excellent for values between 0.9-1. Fig. 5 shows ROC-AUC curves for various ensemble techniques applied in the experimentation. Accuracy and AUC are two important metrics for the binary classification problem. The values of AUC for all the techniques are found to be excellent as per the observation in the ROC curve.

Fig. 6 demonstrates the two cases of the heart disease prediction system as heart disease positive and heart disease negative on a GUI application created using tkinter in Python.

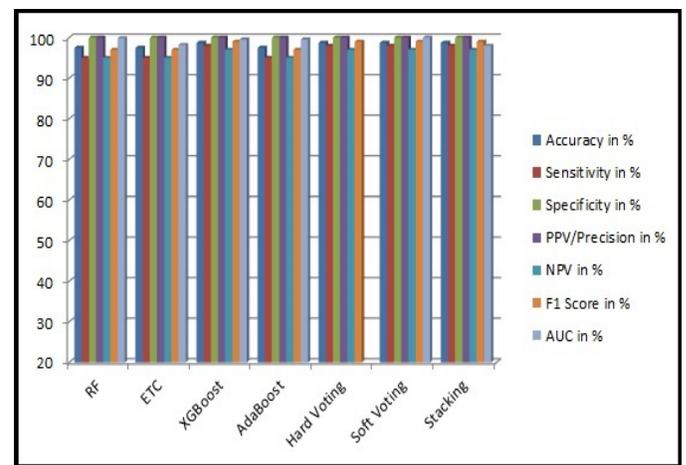


Fig. 4. Performance parameters of proposed ensemble technique for Cleveland dataset.

The proposed work's credibility is demonstrated by comparing its findings for the same dataset with three approaches listed below.

Approach 1: Without Random Over sampling and Isolation Forest for Ensemble Techniques.

Approach 2: Without Isolation Forest with Random Over sampling for Ensemble Techniques.

Approach 3: Without Random Over sampling with Isolation Forest for Ensemble Techniques.

In the first approach, Cleveland dataset is processed without random over sampling and no isolation forest. The results of this implemented methodology are tabulated as shown in Table IV.

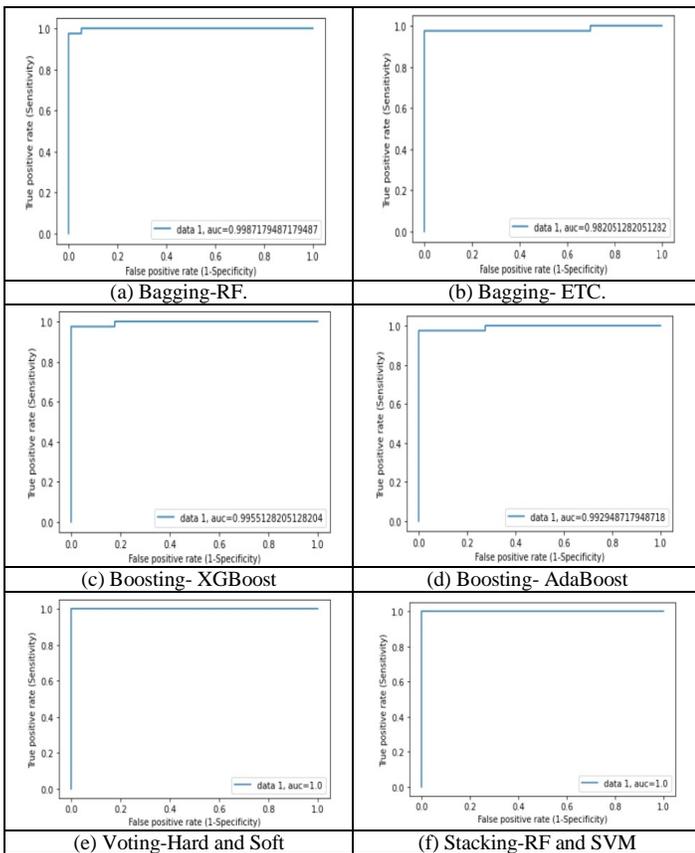


Fig. 5. ROC curve of various ensemble Techniques used in the experiment, displaying corresponding AUC values.

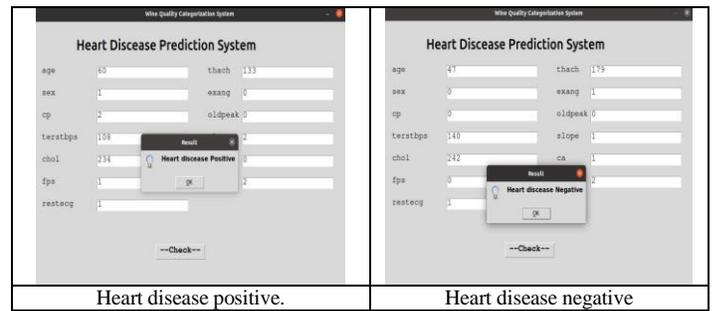


Fig. 6. GUI for Heart disease positive and negative cases using tkinter in python.

The highlighted column in Table IV demonstrates that, soft voting is the best performing ensemble model with 96.05% accuracy for approach 1. But the performance of the proposed research is far better as compared to the implemented methodology in approach 1.

Similarly, in the second approach, Cleveland dataset is processed with random over sampling, no isolation forest and in the third approach Cleveland dataset is processed sampling with isolation Forest, no random over sampling.

The results of these implemented methodologies are tabulated as shown in Table V and Table VI respectively and the best results are highlighted.

The projected results in Tables IV, V and VI reveal that our proposed research work of implementation of Ensemble Techniques with random oversampling and isolation forest give excellent performance in terms of all performance matrices as compared to implementation strategy of all the three approaches.

TABLE IV. PERFORMANCE EVALUATION FOR CLEVELAND DATASET OF 303 INSTANCES WITHOUT RANDOM OVER SAMPLING AND ISOLATION FOREST (APPROACH 1)

Metric	RF	ETC	XGBoost	AdaBoost	Hard Voting	Soft Voting	Stacking
Accuracy	94.73	94.73	94.73	94.73	96.05	96.05	96.05
Sensitivity	92	92	92	92	92	92	92
Specificity	97	97	97	97	100	100	100
PPV	97	97	97	97	100	100	100
NPV	93	93	93	93	93	93	93
F1-Score	0.95	0.95	0.95	0.95	0.96	0.96	0.96
AUC	0.995	0.995	0.995	0.981	--	0.983	0.956
Computational Time in sec.	0.057	0.082	0.154	0.113	0.015	0.011	0.681

TABLE V. PERFORMANCE EVALUATION FOR CLEVELAND DATASET OF 303 INSTANCES WITHOUT ISOLATION FOREST WITH RANDOM OVER SAMPLING (APPROACH 2)

Metric	RF	ETC	XGBoost	AdaBoost	Hard Voting	Soft Voting	Stacking
Accuracy	95.18	95.18	95.18	95.18	95.18	95.18	95.18
Sensitivity	91	91	91	91	91	91	91
Specificity	100	100	100	100	100	100	100
PPV	100	100	100	100	100	100	100
NPV	90	90	90	90	90	90	90
F1-Score	0.95	0.95	0.95	0.95	0.95	0.95	0.95
AUC	0.966	0.947	0.980	0.950	--	0.971	0.971
Computational Time in sec.	0.067	0.071	0.173	0.123	0.024	0.015	0.719

TABLE VI. PERFORMANCE EVALUATION FOR CLEVELAND DATASET OF 303 INSTANCES WITHOUT RANDOM OVER SAMPLING WITH ISOLATION FOREST (APPROACH 3)

Metric	RF	ETC	XGBoost	AdaBoost	Hard Voting	Soft Voting	Stacking
Accuracy	94.44	94.44	94.44	94.44	94.44	94.44	94.44
Sensitivity	89	89	89	89	89	89	89
Specificity	100	100	100	100	100	100	100
PPV	100	100	100	100	100	100	100
NPV	89	89	89	89	89	89	89
F1-Score	0.94	0.94	0.94	0.94	0.94	0.94	0.94
AUC	0.966	0.941	0.966	0.941	--	0.969	0.951
Computational Time in sec.	0.062	0.077	0.140	0.093	0.015	0.0	0.625

## VI. COMPARISON AND DISCUSSION

The reliability of the proposed work is demonstrated by comparing its findings to those of other state-of-the-art current systems conducting imbalanced data processing and outlier detection for the Cleveland dataset as shown in Table VII. Researcher [17] implemented imputation of missing values and outlier removing processes in order to provide quality data to ML model. Mahalanobis distance metric is used to drop the outliers in the data. Further, NB optimized with grid search found to provide accuracy of 84.8%. Instead of conventional model, an ensemble based majority voting scheme is proposed by researchers [18]. Outliers in the dataset are identified and removed using filter based techniques. From different combinations of ensemble, SVM +NB+ MLP ensemble provided highest accuracy of 84%. Researchers [19] proposed machine learning framework using data imbalance technique SMOTE and feature selection technique on ensemble (LR+KNN) classifier for heart disease prediction. Box plot technique is used to identify the outliers. The suggested architecture was assessed on the Framingham, heart disease, and Cleveland dataset and found to outperform them all. SMOTE based ANN [20] is proposed to Cleveland UCI dataset. The imbalance nature of the presented dataset is analyzed, and SMOTE oversampling strategy is proposed to improve the performance of the ANN classifier. Deep learning has been effectively used in heart disease prediction. Researchers experimented isolation forest for outlier detection

for multivariate data using selected features from 13 features of Cleveland dataset [21]. Accuracy is found to be improved but sensitivity and specificity is very poor. Researchers used filter based feature selection and isolation forest for anomalies detection. The proposed approach found to be performing effectively for KNN with eight neighbours on UCI Cleveland dataset [22]. Researchers investigated anomaly detection using K-means clustering algorithm [23]. After removing anomalies, five classification techniques KNN, RF, SVM, NB and LR are used to build the prediction model. It is found that without anomalies RF and NB are performing better as compared to with anomalies in the dataset. DBSCAN is implemented to identify and remove the outliers, a hybrid SMOTE-Edited Nearest Neighbor (SMOTE-ENN) to balance the training data distribution and XGBoost as a classification model for heart disease prediction [24]. Most of the existing research work, proposed to use feature selection extensively. The state of the art research rarely used combination of data balancing and outlier detection in data pre-processing.

The suggested methodology outperforms prior research work utilizing data balancing techniques and outlier identification techniques. The performance of the data balancing solution using Random Oversampling and outlier detection using Isolation Forest on Ensemble Classifier is excellent in all performance metrics. The proposed model in our research outperformed previous models and research findings, with an accuracy of 98.73 % for Cleveland dataset.

TABLE VII. COMPARISON OF PROPOSED METHODOLOGY WITH STATE-OF-THE-ART RESEARCH FOR UCI CLEVELAND DATASET

Author Name, Year, Reference	Methodology	Accuracy	Sensitivity	Specificity	PPV	NPV	F Score	AUC Score
Sivaraman et al. (2021) [17]	Mahalanobis distance+ Grid search optimization +NB	84.8	82.7	86.4	83.2	--	82.8	91
Bashir et al. (2021) [18]	Filter based outlier removing + Majority voting	84	84.80	83.22	--	--	84	--
Rahim et al. (2021) [19]	SMOTE+ Box plot +Feature importance (5 features)+ LR-KNN	98.0	--	--	--	--	--	--
Waqar et al. (2021) [20]	SMOTE+ ANN	96	95.7	--	96.1	--	95.7	100
Bharti et al. (2021) [21]	Isolation Forest+ Lasso FS+DL	94.2	82.3	83.1	--	--	--	--
Ramesh et al. (2022) [22]	Isolation Forest+Filter FS (7 features) +KNN	94.1	94.8	--	91.7	--	90.8	79.9
Ripen et al. (2021) [23]	K means clustering +RF	88	--	87	87	--	--	--
Fitriyani et al. (2020) [24]	DBSCAN+ SMOTE ENN+ XGBoost	98.4	98.3	98.3	98.5	--	98.3	--
Proposed Methodology	Random Oversampling+ Isolation Forest+ Ensemble Techniques	98.73	98	100	100	97	99	100

For binary classification problem of presence and absence of heart disease, accuracy and AUC are the most important metrics. The highlighted row in the Table VII demonstrates that the proposed methodology has highest accuracy and AUC as compared to state of the art research.

## VII. CONCLUSION AND FUTURE RESEARCH

In this paper, we propose ensemble techniques that are supported by Random Oversampling and Isolation Forest for efficiently predicting heart disease. All the ensemble models are found to be performing excellently in all evaluation results. The accuracy range for the Cleveland dataset, for all models is 97.43% to 98.73%, sensitivity 95% to 98%, specificity 100%, precision 100%, NPV 95% to 97 %, F score 0.97 to 0.99, AUC score 0.98 to 1 with less computational overhead.

Most of the previous researchers implemented feature selection techniques to improve the accuracy of ML and DL models. Here we propose a model without any feature selection and achieve remarkably improved performance metrics. Experimental results prove that the ensemble approach with data pre-processing techniques resolves the issue of computational intricacy. Random oversampling and isolation forest significantly improve the performance of ensemble classifiers.

Data balancing with oversampling helps to make data more reliable by avoiding over fitting or under fitting the model. Noisy data removal with an isolation forest improves the quality of the data. The validity of the suggested framework on the UCI Cleveland dataset demonstrates that our framework is both trustworthy and efficient. It incorporates novel pre-processing techniques while also employing an inventive ensemble. Furthermore, the computational time is remarkably reduced with highly reliable results.

In the future, optimization techniques can be implemented for hyper parameter tuning to deploy the model. Because of the NP-hardness of feature selection approaches, a meta-heuristic feature selection method can be devised. There is a strong need for real-world clinical factors that are easily approachable and computed in real-time for the future of clinical cardiac disease detection via ML-centered systems.

## REFERENCES

- [1] S. Vinayaka and P.K. Gupta, "Heart disease prediction systems using classification algorithms," *Proceedings of International conference on Advances in Computing and Data Sciences*, vol. 1244, pp.395-404, July 2020.
- [2] H. Jindal, S. Agrawal, R. Khera et al., "Heart disease prediction using machine learning algorithms," *IOP Conference Series: Materials Science and Engineering*, vol.1022(012072), pp.1-10, 2021.
- [3] I.D.Mienye, Y. Sun, and Z. Wang Z., "An improved ensemble learning approach for the prediction of heart disease risk," *Informatics in Medicine Unlocked*, vol.20(100402), pp.1-5, 2020.
- [4] R.Atallah and A. Al-Mous, "Heart disease detection using machine learning majority voting ensemble method," *Proceedings of the 2nd International Conference on new Trends in Computing Sciences (ICTCS)*, pp.1-6, 2019.
- [5] A.Javeed, S. Zhou, L. Yongjian et al. 4, "A. An Intelligent Learning System Based on Random Search Algorithm and Optimized Random Forest Model for Improved Heart Disease Detection," *IEEE Access*, vol. 27, pp.180235-180243, 2019.
- [6] Y.Muhammad, M.Tahir,M. Hayat M et al., "Early and accurate detection and diagnosis of heart disease using intelligent computational model," *Scientific Reports, nature research*, vol 10(19747), pp.1-17, 2020.
- [7] K.Dissanayake and M.G. Johar, "Comparative Study on Heart Disease Prediction Using Feature Selection Techniques on Classification Algorithms," *Applied Computational Intelligence and Soft Computing*, vol.5581806,pp.1-17, 2021.
- [8] H.Koshimizu, H. Kojima and Y. Okuno Y, "Future possibilities for artificial intelligence in the practical management of hypertension," *Hypertension Research*, vol. 43(12), pp. 1327-1337, 2020.
- [9] M.Rong M, D. Gong and X. Gao , "Feature Selection and Its Use in Big Data: Challenges, Methods, and Trends," *IEEE Access*, vol. 27, pp. 19709-19725, 2019.
- [10] D.Yewale,S.P. Vijayaragavan and M. Munot M, "Decision support system for reliable prediction of heart disease prediction using machine learning techniques: an exhaustive survey and future directions," *International Journal of Engineering Trends and Technology*, vol. 7(4), pp. 316-331, 2022.
- [11] R.C. Ripan,I.H. Sarker, M.H. Furhad, M.M. Anwarand M.M. Hoque, "An Effective Heart Disease Prediction Model Based on Machine Learning Techniques," *Hybrid Intelligent Systems Advances in Intelligent Systems and Computing*, preprints 2020, pp.280-288, 2020.
- [12] H. Kang , "The prevention and handling of the missing data" *Korean Journal of Anesthesiology*, vol. 64(5), pp. 402-406, 2013.
- [13] M.M.Ahsan, M.A.P. Mahmud, P.K. Saha et al., "Effect of Data Scaling Methods on Machine Learning Algorithms and Model Performance," *Technologies*, vol. 9(52), pp. 1-17, 2021.
- [14] B. Krawczyk, "Learning from imbalanced data: open challenges and future directions," *Progress in Artificial Intelligence*, vol. 5, pp.221-232, 2016.
- [15] I.D. Apostolopoulos, "Investigating the Synthetic Minority class Oversampling Technique (SMOTE) on an imbalanced cardiovascular disease (CVD) dataset," *International Journal of Engineering Applied Sciences and Technology*, vol. 4(2020), pp. 431-434, 2020.
- [16] F.T.Liu, K.M. Ting and Z.H. Zhou, "Isolation-based Anomaly Detection," *ACM Transactions on Knowledge Discovery from Data* vol. 6(1), pp.1-39, 2012.
- [17] K. Sivaraman and V. Khanna V, "Machine Learning Models for Prediction of Cardiovascular Diseases," *Journal of Physics: Conference Series*, 2040 012051, 2021.
- [18] S. Bashir, A.A. Almazroi, S. Ashfaq et al., "A knowledge-based clinical decision support system utilizing an intelligent ensemble voting scheme for improved cardiovascular disease prediction," *IEEE Access*, vol. 9, pp. 130805-130822, 2021.
- [19] A. Rahim, Y. Rasheed, F. Azam et al., "An Integrated Machine Learning Framework for Effective Prediction of Cardiovascular Diseases," *IEEE Access*, vol. 9, pp.106575-106588, 2021.
- [20] M. Waqar, H. Dawood, H. Dawood et al., "An Efficient SMOTE-Based Deep Learning Model for Heart Attack Prediction," *Scientific Programming*, vol. 2021(6621622), pp.1-12, 2021.
- [21] R. Bharti, A. Khamparia, M. Shabaz et al., "Prediction of Heart Disease Using a Combination of Machine Learning and Deep Learning," *Computational Intelligence and Neuroscience*, vol. 2021(8387680), pp.1-11, 2021.
- [22] R. TR, U.K. Lilhore, M. Poongodi et al., "Predictive Analysis of Heart Diseases with Machine Learning Approaches," *Malaysian Journal of Computer Science*, vol. 1, pp.132-148, 2022.
- [23] R.C. Ripan,I.H. Sarker,M.H.Furhad et al., "A Data-Driven Heart Disease Prediction Model through K-Means Clustering-Based Anomaly Detection," *SN Computer Science*, vol. 2(112), pp.1-12, 2021.
- [24] N.L.Fitriyani,M. Syafrudin, G. Alfian et al., "HDPm: An Effective Heart Disease Prediction Model for a Clinical Decision Support System," *IEEE Access*, vol. 8(2020), pp.133034-133050,2020.

# Explaining the Outputs of Convolutional Neural Network - Recurrent Neural Network (CNN-RNN) based Apparent Personality Detection Models using the Class Activation Maps

WMKS Ilmini<sup>1</sup>, TGI Fernando<sup>2</sup>

Faculty of Graduate Studies, University of Sri Jayewardenepura, Nugegoda, Sri Lanka<sup>1</sup>  
Intelligent Research Laboratory-Faculty of Computing, General Sir John Kotelawala Defence University,  
Rathmalana, Sri Lanka<sup>1</sup>

Department of Computer Science-Faculty of Applied Sciences,  
University of Sri Jayewardenepura, Nugegoda, Sri Lanka<sup>2</sup>

**Abstract**—This study aims to use the Class Activation Map (CAM) visualisation technique to understand the outputs of apparent personality detection models based on a combination of Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN). The ChaLearn Looking at People First Impression (CVPR'17) dataset is used for experimentation in this study. The dataset consists of short video clips labelled with the Big Five personality traits. Two deep learning models were designed to predict apparent personality with VGG19 and ResNet152 as base models. Then the models were trained using the raw frames extracted from the videos. The highest accurate models from each architecture were chosen for feature visualisation. The test dataset of the CVPR'17 dataset is used for feature visualisation. To identify the feature's contribution to the network's output, the CAM XAI technique was applied to the test dataset and calculated the heatmap. Next, the bitwise intersection between the heatmap and background removed frames was measured to identify how much features from the human body (including facial and non-facial data) affected the network output. The findings revealed that nearly 35%-40% of human data contributed to the output of both models. Additionally, after analysing the heatmap with high-intensity pixels, the ResNet152 model was found to identify more human-related data than the VGG19 model, achieving scores of 46%-51%. The two models have different behaviour in identifying the key features which influence the output of the models based on the input.

**Keywords**—Apparent personality detection (APD); convolutional neural network based recurrent neural network (CNN-RNN); class activation map (CAM); explainable AI (XAI)

## I. INTRODUCTION

Explainable AI (XAI) has gained attention in machine learning as it is crucial to comprehend the behaviour of these models, that is, how these models generate their outputs. These techniques can be used to gain a deeper understanding of the inner workings of a model and can help improve the trust and adoption of AI systems in various applications. Artificial neural networks and deep learning methods are often considered as black boxes, as the inner workings of these techniques and how they produce output based on input are not fully understood. Therefore, researchers tend to explore

techniques to make these into glass boxes, that is, to understand how input features contribute to the output. According to the literature [1], these techniques are divided into different categories.

### • Local Vs Global

- **Local:** Explain how a model makes a single prediction and evaluates its performance on a given set of examples.
- **Global:** Global techniques for explaining a model do not require a specific set of example data and instead provide an overall understanding of how the model works. These techniques can include analysing the model's architecture and studying the relationships between the model's parameters.

### • Model Specific Vs Model Agnostic

- **Model Specific:** These techniques only apply to a single model or a group of specific models.
- **Model Agnostic:** These techniques can be applied to any model to explain the model's predictions.

XAI techniques are grouped into different clusters based on the type of data, the purpose of interpretability, and the flow of interpretation signals, in addition to the two primary categories mentioned above.

The Saliency map is the oldest and most commonly used technique to explain convolutional neural network (CNN) predictions. The saliency map specifies the pixels that activate a particular layer in the network. The literature discloses three main approaches: Deconvolutional Network [2], Backpropagation [3], and Guided Backpropagation [4]. Table I summarises the most popular XAI techniques.

Researchers discovered various XAI methods to understand the deep learning model predictions in addition to the methods mentioned in Table I.

Apparent Personality Detection (APD) based on a person's appearance is a trending research topic in affective computing because apparent personality is helpful in various applications. A few of those applications are listed below:

- **Job Screening:** From the past [10] to the present [11], [12], psychological researchers have tended to find a relationship between personality and job performance. Barrick et al. [11] identified a relationship between Extraversion and Conscientiousness personality traits in the ratings of sales representatives. Inceoglu and Warr [13] conducted a study to reveal the relationships between job engagement and personality. They concluded that there is a relationship between Extraversion, Conscientiousness and Emotional Stability. Hence, the different personality traits contribute to job roles, performance, and satisfaction. Such as, a team leader should have a high level of Extraversion and Conscientiousness and a low level of Neuroticism.
- **Recommendation Systems:** Dhelim et al. [14] discuss the need for personality-aware recommendation systems. Hence, people with the same characteristics act in the same way. It is easy to recommend products or solutions if the customer's personality is known. The authors also mentioned that personality-aware recommendation systems are better when dealing with cold start and data sparsity issues than traditional recommendation techniques.
- **Social Robotics:** A study by Lee et al. [15] found that if the robot's personality is similar to the user's personality, users enjoyed dealing with the robot. Kirby et al. [16] highlight the importance of affective-social robots with emotions and apparent personalities. The robot can identify the user's state and act accordingly. It is essential to consider this when designing social robotics.
- **Personal Assistants:** There are many personal assistants available nowadays, including Apple Siri, Microsoft Cortana, Google Assistant, and Huawei Celia. These personal assistants can be enhanced by adding the automatic personality detection feature, which leads to higher user interaction with personal assistants.
- **Animation Movies:** Designing an animation-movie character is challenging since it should reflect the character's qualities, including personality [17]. Identification of the facial features which contribute to the different personality traits will be beneficial in this field to improve the outcomes.
- **Health Care and Counselling:** In psychology, researchers are researching the relationship between personality and mental health, personality and physical

health and personality and illness. Smith and MacKenzie [18] discuss how personality traits (such as neuroticism) affect a human's health. Hence psychology research proved that our mental and physical health is affected by personality. It will be beneficial to identify personality for early treatment processes and personalised counselling plans based on the personality.

- **Criminology:** Reid [19] explained the connection between personality and crime. Hence, with better personality prediction solutions, authorities can identify and prevent criminal activities.
- **Education and Personalised Learning:** Salazar et al. [20] highlight the importance of having an affective recommendations system in the education field. Moreover, they mentioned that it is vital to change the content based on the learning style, emotions and personality.

According to the review study conducted by [21], psychological studies, political forecasting, forensic, and word polarity detection can also be enhanced by automatic personality detection.

Thus, an individual's apparent personality can be used in different domains to improve performance and effectiveness. Researchers introduced deep learning solutions, including convolutional neural networks and recurrent neural network architectures, to measure apparent personality. After achieving higher accurate predictions by APD deep learning models, researchers tend to find how these models produce the output for given input features using XAI techniques. The other purpose of applying XAI in APD models is to identify prominent facial and non-facial features that affect the personality, which is more important to improve the trust and adoption of AI systems in the above mentioned applications. All works performed in this area used ChaLearn Looking At People First Impression V2 (CVPR'17) dataset [22]. This is the only dataset publicly available with labelled Big Five Personality traits [23].

Zhang and co-workers [24] applied a heatmap feature visualisation technique to visualise the features affecting the APD. They have used different deep learning architectures such as ResNet, DAN, and DAN+. Their study results convey that different models focused on different features of the face, including facial and non-facial data, including background data. Ventura and co-workers [25] conducted a quantitative study to identify prominent facial features and emotions that influence APD.

TABLE I. MOST POPULAR XAI TECHNIQUES APPLIED IN THE FIELD OF CNN

Technique	Year	Local/ Global	Model Specific/ Agnostic	Description
Deconvolutional Network [2]	2013	Local	Specific	Deconvolutional networks work as the inverse of convolution, pooling (unpooling), and inverse of ReLU. This technique recognises the features activated by the immediate layer for the given input. It reconstructs the input from the activations of the layer.
Backpropagation [3]	2014	Local	Specific	For a given input, calculate the gradients concerning the network parameters. This technique highlights the pixel space based on the gradients they receive, which implies the contribution of these pixels to the final output.
Guided Backpropagation [4]	2015	Local	Specific	Guided Backpropagation is a combination of a deconvolutional network and the backpropagation technique. This technique identifies the essential features based on the reconstruction signal's negative values (deconvolutional) and negative values of the input in the forward pass (backpropagation).
CAM [5]	2015	Local	Specific	Class Activation Map (CAM) detects different regions contributing to a given class score. The last fully connected layers are replaced by a global average pooling (GAP) layer, which averages the activations of feature maps. The GAP layer produces a vector, then calculate the weighted sum of the vector's components and sends it to the SoftMax layer. The calculated weighted values help identify the essential features that activate each convolutional feature map by projecting them back.
Deep LIFT [6]	2019	Local	Specific	This technique calculates the activation feature map by multiplying the input with the measured gradients for the given input with a class of interest.
Grad-CAM [7]	2020	Local	Specific	This is a more flexible version than CAM because this produces feature activation with fully connected layers. When the class of interest and input is produced to the network, the network calculates the gradient flow into the final convolutional layer.
Guided Grad-CAM [7]	2020	Local	Specific	Since the Grad-CAM cannot highlight fine-grained regions, the same authors suggest combining the Grad-CAM and Guided Backpropagation techniques to obtain the Guided Grad-CAM.
LIME [8]	2016	Local	Agnostic	Local Interpretable Model-Agnostic Explanations (LIME) manipulate the input data by creating a set of artificial data. These artificial data consist of part of the original input data. The artificial data is then introduced to the model and classified into different categories. Hence, the presence or absence of certain input parts can decide the contribution to the model's output.
SHAP [9]	2017	Local and Global	Agnostic	Shapley Additive Explanations (SHAP) is based on the Shapley values used in game theory. Shapley values are vastly applied in the cooperative game theory to find each player's contribution/ importance. The same theory is applied in the XAI to identify feature importance for the final output.

They applied CAM and Action Unit (AU) [26]. CAM is applied to find the discriminative regions in the scene data. CAM results convey that the facial regions, such as the eye, nose, and mouth areas, contribute to the final prediction. From the Action Coding System, 17 AU was applied to find the influence of emotions in APD. The results indicate that few AUs affected personality detection. They concluded these results with 50 images extracted with the highest personality scores.

Wei et al. [27] applied feature map visualisation to the models they trained to predict the apparent personality. Results show that ResNet identified the facial region as the primary contributor, while DAN and DAN+ activate background data rather than facial data. However, with plain background data, DAN and DAN+ identify facial data, while ResNet fails to identify facial data as primary contributing features. They summarised the model interpretability techniques results with 12 randomly selected images.

Yang and Glaser [28] used saliency map model interpretability techniques to interpret the APD models' outputs. They also concluded that ResNet pre-trained model-based APD architecture could identify facial features as primary contributors. Li et al. [29] calculated heatmap on scene data to identify the most contributing features using the Seaborn Python library [30]. Their findings revealed that critical facial features such as the eye, nose, and mouth

contribute to APD. However, non-facial features such as clothing and furnishing affect the APD model's output. They conducted a quantitative study by considering the face area and heatmap of contributing features and concluded that 73.96% of the highlighted points are face key points (eye, nose, and mouth). They used 32 frames for the experiment from each video from the test dataset of CVPR'17 [22].

A summary of the works conducted in this area used heatmap visualisation techniques such as saliency map techniques to interpret the prediction of APD models. Most of these works concluded that facial regions and non-facial data contribute to the output. A majority of these techniques tend to interpret the CNN architectures' output. These researchers used different pre-trained models in the development and various XAI techniques and concluded that different architectures tend to highlight different areas. Less attention has been paid to work focusing on describing the outputs of Convolutional Neural Networks based Recurrent Neural Network (CNN-RNN) architecture and work on conducting a quantitative study to prove the findings.

#### A. Contribution

Contributions of the work to the APD area are as follows:

- 1) Prior works mainly focused on explaining the CNN-based APD models. This work focused on CNN-RNN models.

2) A quantitative study is conducted to identify primary contributing features for the CNN-RNN-based APD model's output.

The primary aim of this work is to explain the CNN-RNN-based APD models using the CAM technique.

The rest of the paper is organised as follows: Section two discusses the Methodology, Section three contains the Results and Discussion, and Section four contains the Conclusion.

## II. METHODOLOGY

This section includes the methodology followed in this study to explain the predictions of the APD models. Fig. 1 shows the overall methodology followed to identify how the human data (facial and non-facial data excluding background) affected apparent personality.

According to Fig. 1, first, the dataset is pre-processed by dividing it into raw frames. Then the extracted frames were used to train, validate, and test the model. After completing the model development, the CAM visualisation technique was applied to the test dataset. The bitwise intersection between the heatmap and the background removed raw frames was calculated to clarify the facial and non-facial (non-background) features that contributed to the network's output.

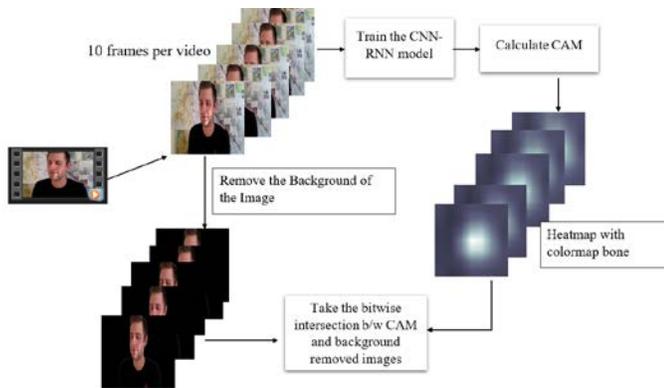


Fig. 1. Overall methodology

### A. Preparation of Data

The experiment used the CVPR'17 [22] dataset, which consists of videos of people facing the camera. These participants are from different nations, ages, and ethnicities. The dataset initially consisted of 3,000 videos which were again processed in 10,000 clips. The training, validation, and test datasets include 6,000, 2,000, and 2,000 video clips. Each video clip is labelled with Big-Five traits ranging from 0 to 1. For model development, ten frames were extracted from each video.

### B. Network Architecture

In CNN-RNN architecture development, the CNN part was developed using pre-trained deep learning models, trained initially on the ImageNet Classification problem (ILSVRC). Two deep learning architectures were designed, developed, and tested to compare the XAI technique findings. VGG19 [31] model is used for the first model, and the second, the

ResNet152 [31] model, is used for the CNN branch. These two models were selected because these are the most common pre-trained models used in several previous works [24], [27], [29]. RNN branch consists of one Gated Recurrent Units (GRU) layer to capture the temporal information, and Fig. 2 illustrates the network's architecture.

### C. Network Parameters

Following are the network parameters used in the current study, finalised after a few experiments conducted with the dataset.

- Batch size = 4
- Early Stop counter: 20
- Maximum number of epochs = 200
- Optimiser: Adam
- Learning rate =  $1 \times e^{-5}$
- Loss: Mean Absolute Error

All experiments were conducted on a precision server with Nvidia RTX 3090 24 GB.

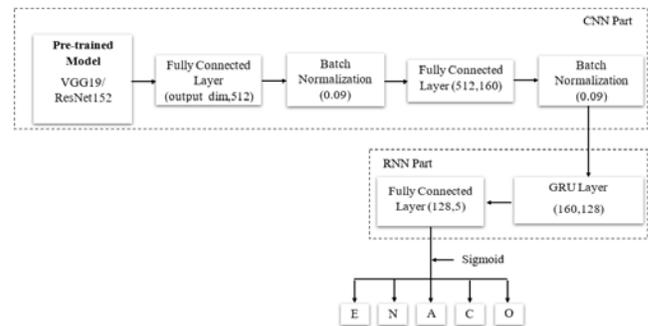


Fig. 2. Deep CNN-RNN network architecture

### D. Visualisation

This study followed the following steps to determine which features (human or background) mainly affected personality prediction in the CNN-RNN deep learning models.

**Step 1:** Removed background data from the raw frames extracted from the video clips (10 frames for each video). Python library Rembg [32] was used to detect human beings from the raw frames. This library uses U2-Net [33] deep learning architecture to detect an object.

$P_{human}$  = Pixels that correspond to the human detected from the raw frames

**Step 2:** Calculated the bitwise intersection between heatmaps and the output of step 1.

Instead of using COLORMAP\_JET [34], the most popular colour map for feature importance visualisation, we used COLORMAP\_BONE [34]. COLORMAP\_BONE, as seen in Fig. 3, uses black and white to represent low and high intensities in pixels. Moderate intensities receive grey colour (in between black and white). Thus, it is more convenient to

identify which features affect more to the output with different intensities.

$$I_1 = P_{human} \cap P_{heatmap_1} \quad (1)$$

$P_{heatmap_1}$  = Pixels which were highlighted by the CAM visualisation technique

$$I_2 = P_{human} \cap P_{heatmap_2} \quad (2)$$

$P_{heatmap_2}$  =  $P_{heatmap}$  pixel values (R, G, B) greater than or equal to 100 (higher intensities)

**Step 3:** Calculated fractions  $f_1$  and  $f_2$ :

$f_1$ : Pixels highlighted by CAM and belongs to the area where the human being exists in the frame / the pixels highlighted by CAM

$f_2$ : Pixels highlighted by CAM with high intensity and belongs to the area where the human being exists in the frame / the pixels highlighted by CAM with high intensity

$$f_1 = \frac{I_1}{P_{heatmap_1}} \times 100 \% \quad (3)$$

$$f_2 = \frac{I_2}{P_{heatmap_2}} \times 100 \% \quad (4)$$

**Step 4:** Followed the above steps for all video files in the test dataset; Measured the average of  $f_1$  and  $f_2$ .

$$F_1 = \frac{\sum_{i=0}^n f_1}{n} \quad (5)$$

$$F_2 = \frac{\sum_{i=0}^n f_2}{n} \quad (6)$$

where  $n = 2000$  (size of the test dataset)

**Step 5:** Repeated the same process for all personality traits.



Fig. 3. Opencv colormaps

### III. RESULTS AND DISCUSSION

The models were trained ten times, and the highest accurate model was selected for feature visualisation. Table II summarises the highest accuracy of each model (VGG19-based CNN-RNN model) and Table III (ResNet152-based CNN-RNN model). The accuracy of the model is calculated using the following equation:

$$\text{Accuracy} = 1 - \frac{1}{N} \sum_{i=1}^N |\text{target}_{ij} - \text{output}_{ij}| \quad (7)$$

$N$ = number of videos, the target is the respective ground-truth value, and output is the predicted value from the model for a given video.

TABLE II. VGG19-BASED MODEL ACCURACY

Big Five Personality Trait	Accuracy
E	90.23%
N	89.95%
A	90.52%
C	90.04%
O	90.16%
Mean Accuracy	90.18%

TABLE III. RESNET152-BASED MODEL ACCURACY

Big Five Personality Trait	Accuracy
E	90.53%
N	90.20%
A	90.21%
C	91.15%
O	90.42%
Mean Accuracy	90.50%

The ResNet152-based model outperforms the VGG19-based model by achieving approximately 90% accuracy for all the traits. While VGG19 based model achieved more than 90% accuracy for all the traits except for neuroticism.

#### A. Visualisation Techniques Results

As mentioned in the methodology section, we calculated the  $F_1$  and  $F_2$  values for all five personality traits with two architectures.

TABLE IV. VGG-19 BASED MODEL F SCORES

Big Five Personality Trait	$F_1$ score	$F_2$ score
E	35.36%	35.76%
N	36.19%	35.40%
A	36.19%	35.40%
C	36.19%	35.40%
O	36.19%	54.00%

Table IV conveys that human data (excluding background) affect personality prediction by 35% - 36%, with  $F_1$  score and  $F_2$  score, except for the Openness trait, which is 54% for  $F_2$ . Hence, with VGG19 openness trait is more influenced by human data with high intensity.

Table V conveys that nearly 35% - 38% ( $F_1$ ) of the facial and non-facial data (excluding background) affected the personality prediction. Furthermore, with  $F_2$  scores, it is 36% to 51%. Since the  $F_2$  scores were calculated using pixels with high intensities in the heatmap, and we can conclude that ResNet152 identified more human data than VGG19 for Extraversion, Openness, Neuroticism, and Conscientiousness traits. In both architectures, the Agreeableness trait is more

affected by background data than other traits (Tables IV and V). In the ResNet152-based model, Extraversion and Openness traits were less affected by background information than other traits (Table V).

TABLE V. RESNET152-BASED MODEL F SCORE VALUE

Big Five Personality Trait	F score	F2 score
E	38.19%	51.44%
N	36.25%	46.95%
A	32.52%	36.08%
C	35.58%	44.14%
O	37.37%	47.68%

Tables IV and V express that the human data (facial and non-background data) affected the personality prediction by nearly 40%, implying that the image's background affected the apparent personality by almost 60%, with  $F_1$  scores. Nevertheless,  $F_2$  scores confirm that high intensities (heatmap) were allocated to human data with ResNet152. The previous works to explain the outputs of the CNN-based APD models also concluded that facial, non-facial, and background data affected the prediction. Zhang and co-workers [24] concluded that different features affect different CNN models designed to predict the apparent personality. As per their demonstration, background data are highlighted as features contributing more to the model prediction. Wei and co-workers [27] also concluded that different models highlighted different features. The models they designed using ResNet and VGGFace-based architectures highlighted different image regions of the input image. Also, they concluded that VGGFace-based architecture is more prone to background data. The current work's quantitative results also convey that models identified different features from the input. Furthermore, ResNet152 is more towards human data rather than the background.

#### IV. CONCLUSION

The primary goal of the current study is to explain the output of the CNN-RNN-based APD models using the CAM as the XAI technique. The results convinced that the models' output is based on the background rather than non-background data (human data, including facial and non-facial data). Usually, the human data (facial and non-facial data excluding background) affects the personality prediction more than the background. However, the findings imply a different conclusion. Even past researchers highlighted this fact with various XAI techniques for CNN-based APD. Hence, the current study with the CNN-RNN APD model also concludes that the background is more influential for APD than human data with the CAM visualisation technique. Also, the models acted differently in the current study because they produced different  $F_1$  and  $F_2$  scores. Furthermore, for Extraversion, Openness, Neuroticism, and Conscientiousness ResNet152 based CNN-RNN models recorded higher  $F_2$  values than  $F_1$ , which implies that more contributing features are from human data. The study's conclusions are derived from an assessment of the deep learning architectures employed and the efficacy of the background removal procedure.

#### ACKNOWLEDGMENT

Our special thanks to the Faculty of Computing of General Sir John Kotelawala Defence University for providing access to servers to experiment.

#### REFERENCES

- [1] P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis, "Explainable AI: A Review of Machine Learning Interpretability Methods," *Entropy*, vol. 23, no. 1, p. 18, Dec. 2020, doi: 10.3390/e23010018.
- [2] M. D. Zeiler and R. Fergus, "Visualizing and Understanding Convolutional Networks," arXiv:1311.2901 [cs], Nov. 2013, Accessed: Sep. 20, 2020. [Online]. Available: <http://arxiv.org/abs/1311.2901>
- [3] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps," arXiv:1312.6034 [cs], Apr. 2014, Accessed: Sep. 19, 2020. [Online]. Available: <http://arxiv.org/abs/1312.6034>
- [4] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, "Striving for Simplicity: The All Convolutional Net," arXiv:1412.6806 [cs], Apr. 2015, Accessed: Sep. 24, 2020. [Online]. Available: <http://arxiv.org/abs/1412.6806>
- [5] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning Deep Features for Discriminative Localization," arXiv:1512.04150 [cs], Dec. 2015, Accessed: Sep. 24, 2020. [Online]. Available: <http://arxiv.org/abs/1512.04150>
- [6] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning Important Features Through Propagating Activation Differences," arXiv:1704.02685 [cs], Oct. 2019, Accessed: Feb. 04, 2021. [Online]. Available: <http://arxiv.org/abs/1704.02685>
- [7] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization," *Int J Comput Vis*, vol. 128, no. 2, pp. 336–359, Feb. 2020, doi: 10.1007/s11263-019-01228-7.
- [8] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why Should I Trust You?: Explaining the Predictions of Any Classifier," arXiv:1602.04938 [cs, stat], Aug. 2016, Accessed: Jan. 17, 2021. [Online]. Available: <http://arxiv.org/abs/1602.04938>
- [9] S. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," arXiv:1705.07874 [cs, stat], Nov. 2017, Accessed: Feb. 05, 2021. [Online]. Available: <http://arxiv.org/abs/1705.07874>
- [10] D. V. Day and S. B. Silverman, "PERSONALITY AND JOB PERFORMANCE: EVIDENCE OF INCREMENTAL VALIDITY," *Personnel Psychology*, vol. 42, no. 1, pp. 25–36, Mar. 1989, doi: 10.1111/j.1744-6570.1989.tb01549.x.
- [11] M. R. Barrick, G. L. Stewart, and M. Piotrowski, "Personality and job performance: Test of the mediating effects of motivation among sales representatives.," *Journal of Applied Psychology*, vol. 87, no. 1, pp. 43–51, 2002, doi: 10.1037/0021-9010.87.1.43.
- [12] G. M. Hurtz and J. J. Donovan, "Personality and Job Performance: The Big Five Revisited," 2000.
- [13] I. Inceoglu and P. Warr, "Personality and job engagement," *Journal of Personnel Psychology*, vol. 10, pp. 177–181, 2011, doi: 10.1027/1866-5888/a000045.
- [14] S. Dhelim, N. Aung, M. A. Bouras, H. Ning, and E. Cambria, "A survey on personality-aware recommendation systems," *Artif Intell Rev*, vol. 55, no. 3, pp. 2409–2454, Mar. 2022, doi: 10.1007/s10462-021-10063-7.
- [15] K. M. Lee, W. Peng, S.-A. Jin, and C. Yan, "Can Robots Manifest Personality?: An Empirical Test of Personality Recognition, Social Responses, and Social Presence in Human-Robot Interaction," *Journal of Communication*, vol. 56, no. 4, pp. 754–772, Dec. 2006, doi: 10.1111/j.1460-2466.2006.00318.x.
- [16] R. Kirby, J. Forlizzi, and R. Simmons, "Affective social robots," *Robotics and Autonomous Systems*, vol. 58, no. 3, pp. 322–332, Mar. 2010, doi: 10.1016/j.robot.2009.09.015.
- [17] M. S. Juhan and N. Ismail, "Character Design towards Narrative Believability of Boboiboy in the Malaysian Animated Feature Film Boboiboy: The Movie (2016)," in *Social Sciences*, 2016, p. 10.

- [18] T. W. Smith and J. MacKenzie, "Personality and Risk of Physical Illness," *Annual Review of Clinical Psychology*, vol. 2, no. 1, pp. 435–467, 2006, doi: 10.1146/annurev.clinpsy.2.022305.095257.
- [19] J. A. Reid, "Crime and Personality: Personality Theory and Criminality Examined," *Inquiries Journal*, vol. 3, no. 01, 2011, Accessed: Sep. 17, 2020. [Online]. Available: <http://www.inquiriesjournal.com/articles/1690/crime-and-personality-personality-theory-and-criminality-examined>
- [20] C. Salazar, J. Aguilar, J. Monsalve-Pulido, and E. Montoya, "Affective recommender systems in the educational field. A systematic literature review," *Computer Science Review*, vol. 40, p. 100377, May 2021, doi: 10.1016/j.cosrev.2021.100377.
- [21] Y. Mehta, N. Majumder, A. Gelbukh, and E. Cambria, "Recent Trends in Deep Learning Based Personality Detection," arXiv:1908.03628 [cs], Aug. 2019, Accessed: Jun. 05, 2020. [Online]. Available: <http://arxiv.org/abs/1908.03628>
- [22] V. Ponce-López et al., "ChaLearn LAP 2016: First Round Challenge on First Impressions - Dataset and Results," in *Computer Vision – ECCV 2016 Workshops*, Cham, 2016, pp. 400–418. doi: 10.1007/978-3-319-49409-8\_32.
- [23] O. P. John and S. Srivastava, *The Big-Five Trait Taxonomy: History, Measurement, and Theoretical Perspectives*. 1999.
- [24] C.-L. Zhang, H. Zhang, X.-S. Wei, and J. Wu, "Deep Bimodal Regression for Apparent Personality Analysis," in *Computer Vision – ECCV 2016 Workshops*, Cham, 2016, pp. 311–324. doi: 10.1007/978-3-319-49409-8\_25.
- [25] C. Ventura, D. Masip, and A. Lapedriza, "Interpreting CNN Models for Apparent Personality Trait Regression," in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Honolulu, HI, USA, Jul. 2017, pp. 1705–1713. doi: 10.1109/CVPRW.2017.217.
- [26] P. Ekman, W. Friesen, and S. Ancoli, "Facial-Sign-Of-Emotional-Experience.pdf," *Journal of Personality and Social Psychology*, vol. 39, pp. 1125–1134, 1980.
- [27] X.-S. Wei, C.-L. Zhang, H. Zhang, and J. Wu, "Deep Bimodal Regression of Apparent Personality Traits from Short Video Sequences," *IEEE Trans. Affective Comput.*, vol. 9, no. 3, pp. 303–315, Jul. 2018, doi: 10.1109/TAFFC.2017.2762299.
- [28] K. Yang and N. Glaser, "Prediction of Personality First Impressions With Deep Bimodal LSTM," *CS 231N Spring 2017*, p. 10, 2017.
- [29] Y. Li et al., "CR-Net: A Deep Classification-Regression Network for Multimodal Apparent Personality Analysis," *Int J Comput Vis*, vol. 128, no. 12, pp. 2763–2780, Dec. 2020, doi: 10.1007/s11263-020-01309-y.
- [30] M. Waskom, "seaborn: statistical data visualization," *JOSS*, vol. 6, no. 60, p. 3021, Apr. 2021, doi: 10.21105/joss.03021.
- [31] 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, Jun. 2016, pp. 770–778. doi: 10.1109/CVPR.2016.90.
- [32] D. Gatis, "Rembg." Jan. 27, 2022. Accessed: Jan. 27, 2022. [Online]. Available: <https://github.com/danielgatis/rembg>
- [33] X. Qin, Z. Zhang, C. Huang, M. Dehghan, O. R. Zaiane, and M. Jagersand, "U<sup>2</sup>-Net: Going Deeper with Nested U-Structure for Salient Object Detection," *Pattern Recognition*, vol. 106, p. 107404, Oct. 2020, doi: 10.1016/j.patcog.2020.107404.
- [34] "OpenCV: ColorMaps in OpenCV." [https://docs.opencv.org/3.4/d3/d50/group\\_imgproc\\_colormap.html](https://docs.opencv.org/3.4/d3/d50/group_imgproc_colormap.html) (accessed Jan. 15, 2023).

# Landmark Recognition Model for Smart Tourism using Lightweight Deep Learning and Linear Discriminant Analysis

Mohd Norhisham Razali<sup>1</sup>, Enurt Owens Nixon Tony<sup>2</sup>, Ag Asri Ag Ibrahim<sup>3</sup>, Rozita Hanapi<sup>4</sup>, Zamhar Iswandono<sup>5</sup>  
Faculty of Business and Management, Universiti Teknologi MARA Sarawak, Kota Samarahan, Malaysia<sup>1,4</sup>  
Faculty of Computing and Informatics, Universiti Malaysia Sabah, Kota Kinabalu, Malaysia<sup>2,3</sup>  
Higher Colleges of Technology, United Arab Emirates<sup>5</sup>

**Abstract**—Scene recognition algorithm is crucial for landmark recognition model development. Landmark recognition model is one of the main modules in the intelligent tour guide system architecture for the use of smart tourism industry. However, recognizing the tourist landmarks in the public places are challenging due to the common structure and the complexity of scene objects such as building, monuments and parks. Hence, this study proposes a super lightweight and robust landmark recognition model by using the combination of Convolutional Neural Network (CNN) and Linear Discriminant Analysis (LDA) approaches. The landmark recognition model was evaluated by using several pretrained CNN architectures for feature extraction. Then, several feature selections and machine learning algorithms were also evaluated to produce a super lightweight and robust landmark recognition model. The evaluations were performed on UMS landmark dataset and Scene-15 dataset. The results from the experiments have found that the Efficient Net (EFFNET) with CNN classifier are the best feature extraction and classifier. EFFNET-CNN achieved 100% and 94.26% classification accuracy on UMS-Scene and Scene-15 dataset respectively. Moreover, the feature dimensions created by EFFNet are more compact compared to the other features and even have significantly reduced for more than 90% by using Linear Discriminant Analysis (LDA) without jeopardizing classification performance but yet improved its performance.

**Keywords**—Scene recognition; convolutional neural network; smart tourism; feature selections

## I. INTRODUCTION

Scene recognition is a crucial aspect for the development of many software applications such as in the area of intelligent robotics, autonomous driving and intelligent video surveillance. Moreover, scene recognition is the basis component in accomplishing the tasks for any object detection tasks [1]. The basic goal of scene recognition is to label all photos of scenes, whether they are outdoor or indoor, semantically and properly.

The magnificent scenery as well as the beautiful and historical landmarks of certain places has become one of the attraction factors for the tourist to come and visit these places. In this context, a software application that equipped with an intelligent landmark detection based on scene recognition algorithm can be developed to serve certain useful tasks. For instance, a tourist may get useful information and

recommendation based on the detected landmark such as the nearby food attractions, and transportation and accommodation information. Besides, the application may assist the tourist agent while guiding the tourist visiting the attraction places. However, the scene recognition is a challenging task due to the difficulty to distinguish the common structure of the public scene objects such building, monuments, parks, beaches and so on [2]. Scene images also might be captured from different angles which triggered the high intra-class difference problems [3].

Deep learning and transfer learning based classification is the emergence approach in any machine learning tasks [4]–[6]. In scene recognition, the pretrained CNN models by using ResNet50 architecture were adopted [4], [5]. Although the classification accuracy obtained was good (92.17% and 94.4%), ResNet50 produced larger features dimensionality. Therefore, there are lot of studies in the other domain have various of Efficient Net (EFFNET) CNN architectures such as masked face recognition [7], smoke detection [8], chest X-ray scanning [9]–[11] and fake face video detection [12] due to its exceptional classification performance as well as to generate lightweight features.

The key contributions of this paper are the proposed super lightweight Landmark recognition model trained by using Convolutional Neural Network (CNN) to address the challenges of distinguishing the common public structure of landmark scenes. The features extracted by using the pretrained CNN model of EfficientNet (EFFNET) which produced the lightest features as compared to the other CNN models. Afterwards, Linear Discriminant Analysis (LDA) feature selection algorithm has been adopted that has significantly reduced the dimensionality of features without sacrificing classification performance at all and even have improved the classification performance. The recognition model training by using CNN was also very efficient as it required very minimal number epoch to complete and yield the best classification performance.

The remainder of the paper is organized in the following way: Section II provide the previous studies conducted in scene recognition. In Section III, the Methodology is described in more detail. Sections IV presents the experimental results. The conclusions and directions for the future studies are presented in Section V.

## II. RELATED WORKS

Scene recognition is a subset of object recognition and can be treated as classification problem to serve certain purposes. It is a problem to describe the content or the objects that exist in the outdoor or indoor scene images. Scene recognition algorithms have adopted in many areas in computing field such as human computer interaction, robotics, smart surveillance system and autonomous driving [1]. Besides, scene recognition was also studied for tourism industry in assisting tourist or tourist guide to recognize the tourism attractive places or landmarks. There are Monulens [13], a real-time mobile-based landmark recognition, Smart Travelling [14] that used to recognize tourist attraction, nearby events, police stations and hospitals, Augmented Reality (AR) based landmark detection [15] and a system to distinguish large number of landmarks. All the aforementioned applications used the handcrafted features such as Histogram of Oriented Gradients (HOG), Scale Invariant Feature Transform (SIFT) and Bag of Features (BoF), and traditional machine learning approach such as Support Vector Machine (SVM). The recent works of scene recognition have shifted to deep learning-based approaches as tabulated in Table I.

TABLE I. RELATED WORKS OF SCENE RECOGNITION

Authors	Dataset	Techniques	Results
[16]	Places image	ImageNet-Linear SVM	Accuracy - 91.9%
[2]	Landmark database	DEep Local Features (DELf)	Specificity-0.99
[4]	Tourist Attraction Images	ResNet50	Accuracy - 92.17%
[5]	Scene images	ResNet-CNN	Accuracy - 94.4%

The study conducted in [16] established Places dataset to benchmark the performance of scene recognition algorithm which was denser in term of density and diversity of scene images in comparisons to the other scene recognition benchmark datasets such as SUN, Scene-15 and MIT Indoor67. The scene recognition algorithm trained by using Places dataset outperformed the accuracy performance of scene recognition algorithm trained by using ImageNet dataset for all scene recognition benchmark datasets. The evaluations were carried out by using CNN based features and linear SVM as classifier. The problem of high density and diversity of scene images as well as to determine whether the scene images contained landmark objects have been also addressed in [2] study. A metric learning-based approach was proposed in which the CNN is trained by curriculum learning technique and updated version of Center loss to overcome large variations of scene images. On the other hand, the existence of landmark objects in scene images determined by calculating distance between the image embedding vector and one or more centroids per class. Other than landmarks diversity issue, the scene recognition algorithm is also facing the high inter-class similarities where numerous landmarks have very similar building or architecture design. To overcome this problem, the CNN model based on ResNet50 was adopted in [4] to classify tourist attraction places in Jakarta, Indonesia such as Cathedral Church, Jakarta Old Town, Istiqlal Mosque and Maritime

Museum. The ResNet based model also demonstrated exceptional performance in [5] via its proposed method namely Scene-RecNet to classify the aerial scene views such as airports, forests and rivers. The Scene-RecNet was more versatile and stable as the features are adjusted and modified in the convolutional and fully-connected layers that eventually improved the processing speed, small storage space and good recognition accuracy.

Table II shows the summary of previous studies that have adopted deep learning approaches, specifically transfer models.

TABLE II. RELATED WORKS OF DEEP LEARNING

Authors	Dataset	Techniques	Results
[17]	Land images	LeNet+Bagging based CNN	Recall-0.784
[7]	Face Mask	EFFNET based CNN	Accuracy-0.9972
[18]	Computer Tomography (CT) Images	Fusion of a moment invariant (MI) method+ ResNet150+VGG16	Accuracy-0.93
[8]	Smoke detection images	EFFNET based CNN	Accuracy-0.9818
[9]	Chest X-ray	DenseNet+EFFNetB0+Bi-LSTM	Accuracy-92.489%
[10]	Chest X-ray	EfficientNet-B2+CNN	Accuracy-96.33%
[12]	Fake Face Video	EFFNetB5+CNN	Accuracy-74.4%
[11]	Chest X-ray	EFFNetB0+CNN	Accuracy - 95.82%
[19]	TripAdvisor and Google	CNN	Accuracy-46.4%

The study conducted in [17] addressed the problem of land-use classification at the hilly and mountainous area by using ensemble learning approaches to improve the overall classification accuracy performance and classes number optimization to solve classification accuracy problem for coniferous forest. The bagging-based CNN using Bagging (Bootstrap AGGregatING) ensemble classifier is capable to overcome the problem of unstable procedures which means the great impact on classification due to minor differences of the data. Whereby the optimization of the classes' number was carried out by utilizing spectral clustering (SC) that divides data into subsets based on its similarity. The pre-trained LeNet CNN architecture have used for feature extraction. The pre-trained CNN architecture was also proposed in [18] for automatic screening of COVID-19. Specifically, two pre-trained CNN architecture ResNet50 and VGG16 were fused with the combination of Moment Invariant methods that improved the performance of previous COVID-19 classification models. It is also worth to note that many previous studies were adopted variant of EfficientNet (EFFNET) CNN architectures for extracting the features from the X-ray to detect lung related diseases. A variant of EFFNET namely EFFNETB0 with Bi-LSTM was proposed by [9] detect Covid-19 faster and with high accuracy low cost on chest X-ray images. Along with that, the features from EFFNETB0 were fused with DenseNet121 and LAB and CIE color space. The model training was performed by using Bi-LSTM classifier that yield the best classification accuracy as compared to the other ensemble classifiers. Similar techniques

were also used in [11] to detect COVID-19 from lung X-ray. The other variant of EFFNET so called EFFNET B2 was found to be most effective as compared to the other variants in [10] to reduce the class imbalance problem for diagnosing pneumonia from chest X-RAY. The fine tuning on EFFNET architecture provides desirable impacts which reduce computational effort and the use of batteries. The evaluation of several EFFNET variants were also carried out in [12] to detect fake face video in social media website. Based on the evaluation, the optimal performance of detection is by using EFFNET B4 and B5 and the classification accuracy performance drops when using EFFNET B6 and B7. Next, The EFFNET with Linear SVM were used to address the issues images complexity to recognize the face mask wearing in [7]. In this study, the classification accuracy EFFNET has outperformed the other CNN models using DENSENET201, NASNETLARGE and INCEPTIONRESNETV2 with very light size of features. The lightness of features produced by EFFNET have been utilized by [8] through the proposed novel lightweight smoke detection for detecting fire in its early stages. A module for smoke region segmentation was also proposed in this study where the encoder-decoder approach with atrous separable convolutions were investigated.

According to the comprehensive survey conducted by [1], the top three performance recognition approaches fall under Patch Feature Encoding, Discriminative Region Detection and Hybrid Deep Models. Specifically, the CNN based feature extraction using ResNet-152, AlexNet and SE-ResNeXt-101 were recorded the significant performance on Scene-15, Sports-8, Indoor-67 and SUN-397.

Based on the discussions of the previous studies, it can be summarized that the pretrained CNN architecture is flexible and capable to provide robust recognition performance in various fields and domains. The CNN architecture is flexible as the layers and its parameters can be easily fine tuned to fit the requirement of data and optimum performance could be achieved. In particular, the EFFNET based CNN architecture has proven quite decent performance so far in terms of classification performance as well as to produce lightweight features. Therefore, the use EFFNET also might be extended in the domain of scene recognition to overcome the issue of high inter-class similarity in scene images.

### III. METHODOLOGY

This section describes the methodology undertaken to carry out this research, as depicted in Fig. 2. The methodology consists of four parts which are data acquisition, feature extraction, feature selection and model training.

#### A. Experimental Setup

The experiment in this study was performed by using Python libraries based on Spyder 4.2.2 and PyCharm 2020.3.3

(Community Edition) software tools. Specifically, the feature extractions and classifications were performed by using Scikit-learn and Keras libraries.

#### B. Scene Recognition Model Training

The landmark recognition model training consists of four main steps which are data acquisition, feature extraction, feature selection and classification model training.

1) *Data acquisition:* The images for UMS Landmark Dataset were captured with a Nikon D7100 camera with a resolution of  $6000 \times 4000$  pixels between 10.00 am. and 11.00 am. Fig. 1 shows the image samples of the popular landmarks in UMS [20]. This dataset has been made public and is available for download on the Kaggle website [21].

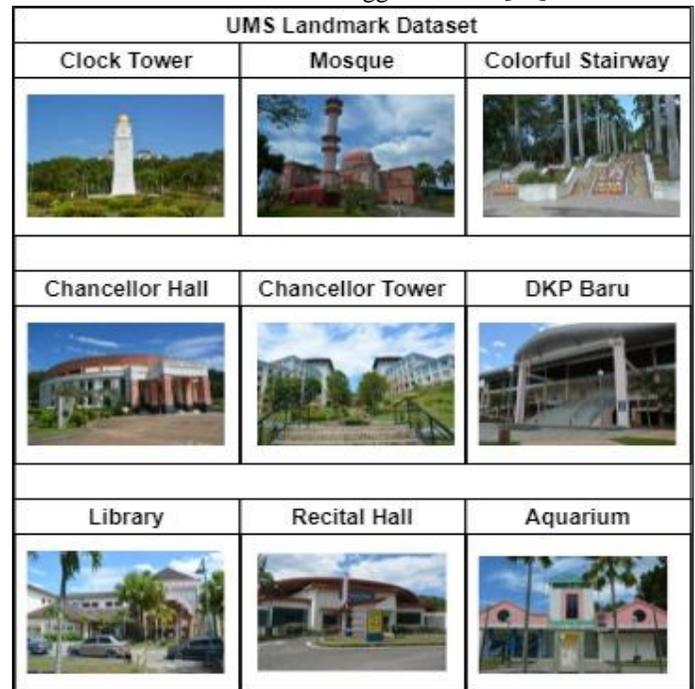


Fig. 1. Samples of UMS landmark dataset

Based on Fig. 1, there are nine categories of landmark consisting around 100 images with different camera angles. These landmarks are the popular tourist attractions for sightseeing and photography. Aside from this dataset, the public Scene-15 dataset [22] for scene recognition benchmarking was also evaluated in order to test the efficacy of the landmark recognition algorithm. This dataset contained 15 scene categories, comprising outdoor and indoor sceneries. There were 200 to 400 images in each category with an average resolution of  $300 \times 250$  pixels.

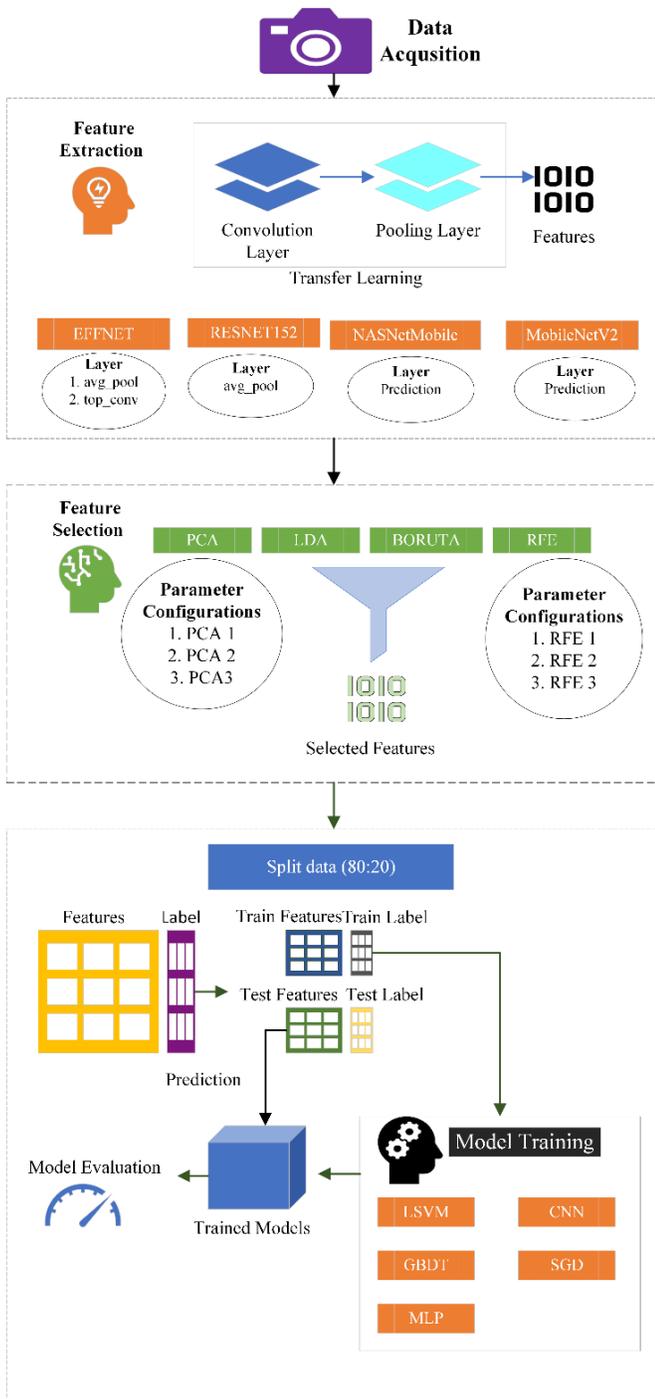


Fig. 2. Methodology

2) *Feature extraction:* Feature extraction is a process to transform the representation of the data into meaningful semantics for determining the category of the data in classification. In this work, the feature extraction was carried out by using transfer learning approach. The features of the images were extracted by re-using the model weights on the pre-trained Convolutional Neural Network (CNN) model. Transfer learning reduces the time it takes to train a neural

network model and lead in decreasing generalization error. The extracted features of an image had created a vector of values that the model would use to characterize the image features. These characteristics were used in designing a new model.

In particular, four pre-trained CNN model were evaluated for feature extraction, which are Efficient Net (EFFNET) [23], RESNET152 [24], NASNetMobile [25] and MobileNetV2 [26]. EFFNET has been adopted and demonstrated to have an outstanding performance in recent studies such as in the Covid-19 detection based on chest X-Ray [9], [27], smoke detection [8], fake video detection [12], pneumonia classification [10], masked face detection [7] and food recognition [28]. Meanwhile, the RESNET152 was also reported to have a good performance for scene recognition [1].

Many previous studies have shown that the NASNetMobile model performs well, such as the classification of rice diseases with an accuracy of 85.9% [29], ECG signal classification for cardiac examination [30] with an accuracy of 97.1 %, lung nodule classification from CT lung images with an accuracy of 88.28% [31] and skin lesion classification from dermoscopic images with an accuracy of 88.28% [32]. For on-device and embedded applications, the proposed MobileNetV2 has a low-latency, low-computation architecture. For instance, MobileNetV2 was used as an embedded food recognizer [33].

The pretrained CNN models were built with various layer types. In this work, two layer types of EFFNET layer were chosen to generate the feature matrices, namely *top\_cov* and *avg\_pool*. The model weights used in the EFFNET were ImageNet and both layers produced 62,720 and 1,280 feature dimensions. On the other hand, the *avg\_pool* was the selected layer to generate 2048 features dimensions for RESNET152 model. Then, both NASNetMobile and MobileNetV2 produced 1000 feature dimensions.

The extracted features consist of one dimensional (1D) features matrix which will be fed into the traditional machine learning classifiers and the 1D CNN classifier (Conv1D). To work with 2D CNN classifier (Conv2D), the 1D features matrix was reshaped into 2D features matrix. The *top\_cov* and *avg\_pool* layers in EFFNET produced (16, 16, 5) and (16,16,245) output shape after being reshaped. Meanwhile, the *avg\_pool* layer of RESNET152 produced (32, 32, 2) feature shape after being reshaped. Meanwhile the *prediction* layer of NASNetMobile and MobileNetV2 generated a (2, 2, 250) feature shape. The feature shape represents the height, width and depth of the images which make the edge and colors of the spatial features to be detected.

3) *Classification model:* The extracted Conv1D or 1D features as described in (2) were fed to Linear Support Vector Machine (LSVM), CNN (1D), Gradient-Boosting Decision Tree (GBDT), Stochastic Gradient Descent (SGD) and Multilayer Perceptron (MLP). Linear kernel is applied and one-versus all (OVA) training strategy is used in LSVM. The parameters used for the classifiers during the experiment are shown in Tables III, IV, V and VI.

TABLE III. LSVM PARAMETERS

Parameters	Value	Description
Penalty	l2	Specifies the norm used in the penalization. The 'l1' leads to coef_ vectors that are sparse.
Loss	square_hinge	Specifies the loss function. 'hinge' is the standard SVM loss (used e.g. by the SVC class) while 'squared_hinge' is the square of the hinge loss.
Dual	1e-4	Tolerance for stopping criteria.
C	1.0	Regularization parameter. The strength of the regularization is inversely proportional to C. Must be strictly positive.
Multi-class	ovr	Determines the multi-class strategy if y contains more than two classes. "ovr" trains n_classes one-vs-rest classifiers, while "crammer_singer" optimizes a joint objective over all classes

TABLE IV. GBDT PARAMETERS

Parameters	Value	Description
Loss	deviance	The loss function to be optimized. 'deviance' refers to deviance (= logistic regression) for classification with probabilistic outputs
learning_rate	0.1	Learning rate shrinks the contribution of each tree by learning_rate. There is a trade-off between learning_rate and n_estimators.
n_estimators	100	The number of boosting stages to perform. Gradient boosting is fairly robust to over-fitting so a large number usually results in better performance.
subsample	1.0	The fraction of samples to be used for fitting the individual base learners. If smaller than 1.0 this results in Stochastic Gradient Boosting. subsample interacts with the parameter n_estimators. Choosing subsample < 1.0 leads to a reduction of variance and an increase in bias.
criterion	friedman_mse	The function to measure the quality of a split

TABLE V. SGD PARAMETERS

Parameters	Value	Description
Loss	hinge	Defaults to 'hinge', which gives a linear SVM
penalty	l2	Defaults to 'l2' which is the standard regularizer for linear SVM models
alpha	0.0001	Constant that multiplies the regularization term
fit_intercept	True	Whether the intercept should be estimated or not. If False, the data is assumed to be already centered.
max_iter	1000	The maximum number of passes over the training data (aka epochs). It only impacts the behavior in the fit method, and not the partial_fit method.

TABLE VI. MLP PARAMETERS

Parameters	Value	Description
hidden_layer_sizes	(100,)	The ith element represents the number of neurons in the ith hidden layer.
activation	relu	Activation function for the hidden layer.
solver	adam	The solver for weight optimization.
alpha	0.0001	0.0001
batch_size	auto	Size of minibatches for stochastic optimizers
learning_rate	constant	Learning rate schedule for weight updates.

On the other hand, the Conv2D training features produced by EFFNET were fed into 2D Convolutional Neural Network classifier which is a fully connected layer. Table VII shows all the layers, the output shapes and the total parameters for EFFNET (*avg\_pool*), EFFNET (*top\_conv*) and RESNET152.

TABLE VII. 2D CONVOLUTIONAL NEURAL NETWORK ARCHITECTURE

Layer (type)	Output Shape			Parameters		
	EFFNET (AvgPool)	EFFNET (TopConv)	RESNET152	EFFNET (AvgPool)	EFFNET (TopConv)	RESNET152
conv2d (Conv2D)	(None, 16, 16, 32)	(None, 16, 16, 32)	(None, 32, 32, 32)	1472	70592	608
dropout (Dropout)	(None, 16, 16, 32)	(None, 16, 16, 32)	(None, 32, 32, 32)	0	0	0
conv2d_1 (Conv2D)	(None, 14, 14, 32)	(None, 14, 14, 32)	(None, 30, 30, 32)	9248	9248	9248
max_pooling_2d (MaxPooling2D)	(None, 7, 7, 32)	(None, 7, 7, 32)	(None, 15, 15, 32)	0	0	0
flatten (Flatten)	(None, 1568)	(None, 1568)	(None, 7200)	0	0	0
dense (Dense)	(None, 512)	(None, 512)	(None, 512)	803328	803328	3686912
dropout_1 (Dropout)	(None, 512)	(None, 512)	(None, 512)	0	0	0
dense_1 (Dense)	(None, 12)	(None, 12)	(None, 12)	6156	6156	6156
Total params	820,204	883,168	3,702,924			
Trainable params	820,204	883,168	3,702,924			
Non-trainable params	0	0	0			

CNN possesses convolution layer that has several filters to perform the convolution operation, which are RELU, pooling layer, and fully connected layer. The RELU layer produces the rectified feature map by performing the operation on the elements. The rectified feature map next feeds into a pooling layer. Pooling is a down-sampling operation that reduces the dimensions of the feature map. The rectified feature map is fed into a pooling layer. Pooling is a down-sampling operation that decreases the feature map's dimensionality. By flattening the

two-dimensional arrays from the pooled feature map, the pooling layer turns them into a single, long, continuous, linear vector. When the flattened matrix from the pooling layer is given as an input, a fully connected layer forms classifiers the images.

The dataset will undergo training and testing phase in creating the classification model. In CNN, the epoch refers to the number of times the model trains all datasets. Whereby, batch size is a small amount of data used for training. A suitable number of epochs needs to be adjusted until a small gap between test and training error can be observed. When the appropriate number of epochs is not chosen, underfitting and overfitting problems occur.

The learning rate determines how frequently the weight in the optimization method is updated. Fixed learning rate is used and the Adam is chosen as optimizer.

Dropout is a better regularization strategy for deep neural networks to avoid overfitting. The method essentially removes units from a neural network based on the probability desired. A default value of 0.5 is set in this experiment. Loss function measure the successfulness of classification and in this experiment by defining the distance between two data points. In this experiment, the categorical cross-entropy loss function was used.

4) *Feature selection*: Feature selection plays important roles to improve the performance of recognition model by reducing the features dimensionality and transforming the feature into meaningful features [34], [35]. The meaningful features are characterized by the features that are more salient, less overfit and reduced the training execution time which eventually improve the accuracy performance [36]. In this work, Principal Component Analysis (PCA) [37], Linear Discriminant Analysis (LDA) [38], Boruta [39] and Recursive Feature Elimination (RFE)[40] were evaluated.

Table VIII shows the number of features selected after performing the feature selection algorithms. Unlike PCA, LDA and RFE, Boruta provided automatic mechanism in determining the number of features. Therefore, manual parameter configurations to determine the number of features selected were not required. Meanwhile, the number of features selected for LDA need to be set to less or equal to the total class in the dataset. For PCA and RFE, experiments were conducted to test three configurations with different percentages of features selected, which are 70%, 40% and 10%.

TABLE VIII. NUMBER OF FEATURES SELECTED

Feature Selection	Configurations	Features	UMS Dataset	Scene-15 Dataset
PCA	PCA1 (70%)	EFFNET	896	
		RESNET152	819	
		NASNETMobile	700	
		MobileNetV2	700	
	PCA2 (40%)	EFFNET	512	
		RESNET152	205	
		NASNETMobile	400	
		MobileNetV2	400	
	PCA3 (10%)	EFFNET	128	
		NASNETMobile	100	
		MobileNetV2	100	
		LDA	EFFNET	8
RESNET152				
NASNETMobile				
MobileNetV2				
BORUTA	EFFNET	749	372	
	RESNET152	479	149	
	NASNETMobile	677	61	
	MobileNetV2	777	158	
RFE	RFE1 (70%)	EFFNET	896	
		RESNET152	1434	
		NASNETMobile	700	
		MobileNetV2	700	
	RFE2 (40%)	EFFNET	512	
		RESNET152	819	
		NASNETMobile	400	
		MobileNetV2	400	
	RFE2 (10%)	EFFNET	128	
		RESNET152	205	
		NASNETMobile	100	
		MobileNetV2	100	

5) *Classification model performance metrics*: The model's overall performance on the testing set was measured using the accuracy metric as the performance metric. Assume that CM is a n by n confusion matrix, with n equaling the total number of various scene categories. Next, the actual category is indicated by the row of CM, while the anticipated category is indicated by the column of CM. Then, let C (i,j) denote the value of the CM cell in index row I and column j, with i,j=1,2,...,n. The following equation defined the accuracy metrics:

$$accuracy = \frac{\sum_{i,j=1}^n C_{i,j}}{\sum_{i=1}^n \sum_{j=1}^n C_{i,j}} \quad (1)$$

#### IV. FINDINGS

This section presents the analysis from the experiment results comprising feature extraction, classification and feature selection performance. The first part of this section presents the discussion of classification performance evaluation, the second part discusses about the feature dimensions size, shape and the number of epoch used in CNN training, followed by the performance analysis for feature selection.

##### A. Classification Performance

Table IX shows the recognition accuracy of feature extraction based on EFFNET, RESNET152, NASNetMobile and MobileNetV2 and classification by using Linear SVM (LSVM), CNN (2D), CNN (1D), Gradient-Boosting Decision Tree (GBDT), Stochastic Gradient Descent (SGD) and Multilayer Perceptron (MLP) on UMS-Scene and Scene-15 dataset.

TABLE IX. CLASSIFICATION ACCURACY COMPARISONS BETWEEN UMS LANDMARK AND SCENE-15 DATASET

Feature Extraction	Layer Name	Classification	UMS-Scene	Scene-15
EFFNET 1	avg_pool	LSVM	<b>1.00</b>	0.94
		CNN (2D)	<b>1.00</b>	0.85
		CNN (1D)	<b>1.00</b>	<b>0.94</b>
		GBDT	<b>1.00</b>	0.68
		SGD	<b>1.00</b>	0.68
		MLP	0.44	0.43
EFFNET 2	top_conv	LSVM	0.94	<b>0.94</b>
		CNN (2D)	0.95	0.91
		CNN (1D)	<b>1.00</b>	0.92
		GBDT	<b>1.00</b>	0.66
		SGD	<b>1.00</b>	0.92
		MLP	0.12	0.40
RESNet152	avg_pool	LSVM	<b>1.00</b>	<b>0.62</b>
		CNN (2D)	0.85	0.58
		CNN (1D)	<b>1.00</b>	0.62
		GBDT	<b>1.00</b>	0.41

		SGD	0.95	0.37
		MLP	0.12	0.23
NASNetMobile	prediction	LSVM	0.77	0.68
		CNN (2D)	0.13	0.38
		CNN (1D)	<b>1.00</b>	<b>0.74</b>
		GBDT	0.99	0.55
		SGD	0.82	0.58
		MLP	0.33	0.39
MobileNetV2	prediction	LSVM	0.85	0.69
		CNN (2D)	0.13	0.36
		CNN (1D)	<b>1.00</b>	<b>0.82</b>
		GBDT	<b>1.00</b>	0.07
		SGD	0.77	0.68
		MLP	0.56	0.34

In comparison to the Scene-15 dataset, most of the algorithms performed well on the UMS landmark dataset, as shown in Table IX. As the UMS landmark dataset had a higher image resolution, the quality of the collected images was more likely to have influenced the result. The bar charts in Fig. 3, Fig. 4, Fig. 5, and Fig. 6 show how the features and classifiers employed in the UMS landmark and Scene-15 datasets compare in terms of performance. The classification accuracy of various features on various classifiers is shown in Fig.3. EFFNET with avg\_pool layer is the best feature due to its perfect achievement on all classifiers except MLP. To demonstrate its efficacy, Fig. 4 shows the classification accuracy of various classifiers on various features. Except for NASNetMobile, CNN 1D and GBDT had been found to be resilient to a variety of features, including the ability to attain 100% classification accuracy on all features. In contrary, CNN 2D performed poorly with NASNetMobile and MobileNetV2. This was most likely because the 2D shape features generated by the CNN 2D classifier were incompatible.

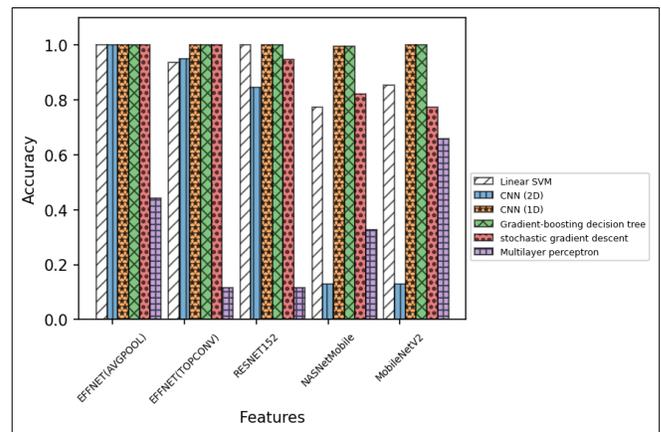


Fig. 3. Performance of features on UMS landmark dataset

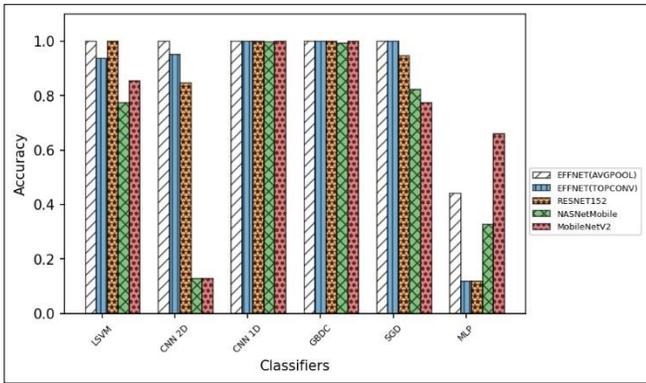


Fig. 4. Performance of classifiers on UMS-landmark dataset

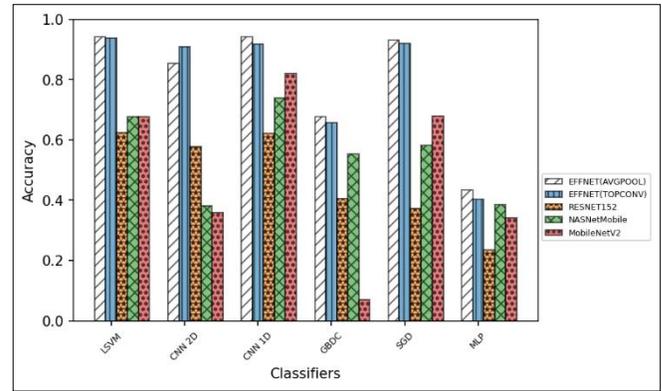


Fig. 6. Performance of classifiers on scene-15 dataset

EFFNET based features performed well across many classifiers in the Scene-15 dataset, apart from GBDC and MLP, as shown in Fig. 5. RESNet152, NASNetMobile, and MobileNetV2, on the other hand, produced less discriminative features. Fig. 6 shows that LSVM and CNN 1D perform consistently across all features and worked exceptionally well with EFFNET features. GBDC and MLP, on the other hand, only achieved 67.61% and 43.39% accuracy, respectively. Moreover, the CNN 2D and SGD only worked well with EFFNET features. Overall, the best classification accuracy on the Scene-15 dataset was 94.26% using CNN 1D classifier and EFFNET (AVGPOOL) features. Based on the study conducted in [1], the RESNet152 indeed yielded the best performance on Scene-15, Sports-8, Indoor-67 and SUN-397. However, based on the result of the experiment in this paper revealed that the EFFNET have better performance on Scene-15 dataset. Next, the confusion matrix of classification accuracy is illustrated in Fig. 7.

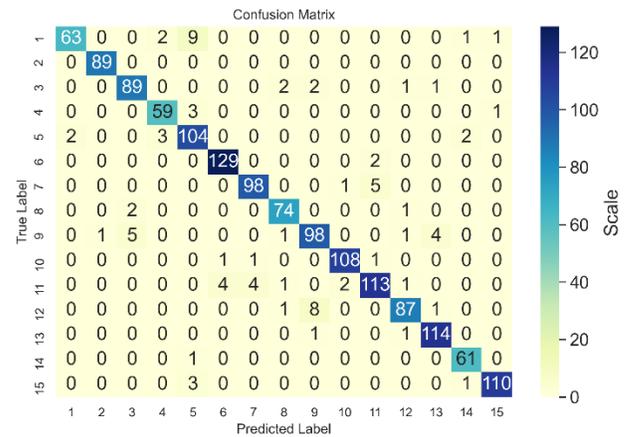


Fig. 7. Confusion matrix of classification using CNN 1D-EFFNET (AVGPOOL) on Scene-15 dataset

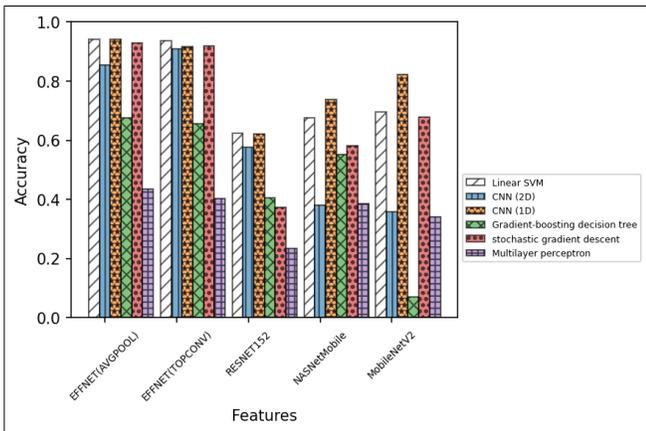


Fig. 5. Performance on features on scene-15 dataset

As plotted in Fig. 7, there are few scene images had been miscategorized. For instance, category 1 (office) was classified as category 5 (store), category 7 (tall building) was classified as category 11 (coast), category 9 (street) was classified as category 3 (living room), category 13 (mountain) was classified as category 9 (open country), and category 12 (open country) could be classified as category 9 (open country) (street). This shows that the high inter-class similarity classification problem still exists due to the appearance diversity of scene photos.

Table X and Table XI shows the precision, recall, F1-score and sup. (support) performance of the algorithms on UMS landmark dataset and Scene-15 dataset. Precision is the capability of a classifier to avoid classifying a negative instance as positive. It is described for each class as the proportion of true positives to the total of true positives and

false positives. Recall is the capacity of a classifier to find all instances that are positive. It is described as the ratio of true positives to the total of true positives and false negatives for each class. A weighted harmonic mean of recall and precision is used to get the F1 score, with the best result being 1.0 and the lowest being 0.0. Due to the inclusion of precision and recall in their computation, F1 scores are lower than accuracy measurements. Support is the number of instances of the class that occur in the particular dataset. The requirement for stratified sampling or rebalancing may be indicated by unbalanced support in the training data, which may point to structural flaws in the classifier's reported scores. Imbalanced support in the training data may indicate structural weaknesses in the reported scores of the classifier and could indicate the need for stratified sampling or rebalancing.

TABLE X. CLASSIFICATION PERFORMANCE ON UMS LANDMARK DATASET

Feature Extraction	Classifier	Prec.	Rec.	F1-Score	Sup.
EFFNet 1	LSVM	1	1	1	309
	CNN (2D)	0.99	0.99	0.99	281
	CNN (1D)	1	1	1	310
	GBDT	1	1	1	281
	SGD	1	1	1	281
	MLP	0.35	0.44	0.36	281
EFFNet 1	LSVM	1	1	1	309
	CNN (2D)	0.96	0.95	0.95	281
	CNN (1D)	1	1	1	310
	GBDT	1	1	1	282
	SGD	1	1	1	282
	MLP	0.01	0.12	0.02	282
RESNet152	LSVM	1	1	1	309
	CNN (2D)	0.86	0.85	0.84	281
	CNN (1D)	1	1	1	310
	GBDT	1	1	1	282
	SGD	0.95	0.95	0.94	282
	MLP	0.01	0.12	0.02	282
NASNetMobile	LSVM	0.82	0.77	0.74	282
	CNN (2D)	0.02	0.13	0.03	281
	CNN (1D)	1	1	1	310
	GBDT	0.99	0.99	0.99	282
	SGD	0.88	0.82	0.79	282
	MLP	0.34	0.33	0.27	282
MobileNetV2	LSVM	0.87	0.85	0.86	282
	CNN (2D)	0.02	0.13	0.03	281
	CNN (1D)	1	1	1	310

	GBDT	1	1	1	282
	SGD	0.88	0.77	0.75	282
	MLP	0.56	0.66	0.59	282

TABLE XI. CLASSIFICATION PERFORMANCE ON SCENE-15 DATASET

Feature Extraction	Classifier	Prec.	Rec.	F1-Score	Sup.
EFFNet 1	LSVM	0.94	0.94	0.94	1480
	CNN (2D)	0.83	0.83	0.83	1167
	CNN (1D)	0.94	0.94	0.94	1481
	GBDT	0.69	0.68	0.67	1346
	SGD	0.69	0.68	0.67	1346
	MLP	0.34	0.43	0.36	1346
EFFNet 2	LSVM	0.94	0.94	0.94	1480
	CNN (2D)	0.91	0.91	0.91	1167
	CNN (1D)	0.92	0.92	0.92	1481
	GBDT	0.68	0.66	0.6	1480
	SGD	0.92	0.91	0.92	1480
	MLP	0.43	0.4	0.5	1480
RESNet152	LSVM	0.63	0.62	0.63	1480
	CNN (2D)	0.56	0.55	0.54	1167
	CNN (1D)	0.63	0.62	0.62	1481
	GBDT	0.42	0.41	0.4	1346
	SGD	0.52	0.37	0.31	1346
	MLP	0.14	0.23	0.17	1346
NASNetMobile	LSVM	0.69	0.68	0.66	1346
	CNN (2D)	0.29	0.38	0.31	1350
	CNN (1D)	0.74	0.74	0.74	1481
	GBDT	0.52	0.55	0.5	1346
	SGD	0.67	0.58	0.57	1346
	MLP	0.25	0.39	0.29	1346
MobileNetV2	LSVM	0.7	0.69	0.69	1346
	CNN (2D)	0.28	0.36	0.3	1350
	CNN (1D)	0.82	0.82	0.82	1481
	GBDT	0.04	0.07	0.05	1346
	SGD	0.72	0.68	0.67	1346
	MLP	0.26	0.34	0.28	1346

### B. Features Shape and Number of Epoch

The extracted features were reshaped into 1D and 2D representations, as can be referred in Table XIII. The 1D feature shape was being fed to LSVM, CNN 1D, GBDT, SGD, and MLP, whereby the 2D feature shape was being fed to CNN 2D classifier. For both datasets, Table XII and Table XIII show the

features' form as well as the best number of epoch for training the CNN. As seen in Table XII, the EFFNET generated the largest 1D features (62720) by using the average pool layer. NASNetMobile and MobileNetV2, on the other hand, generated the smallest number of features (1000). The best classification accuracy can be obtained by using only 30 epochs via CNN 1D for all the features. Whereby, the number of epochs was higher for training the CNN 2D are except for MobileNetV2.

TABLE XII. FEATURES' DIMENSION SIZE AND EPOCH FOR UMS LANDMARK DATASET

Feature Extraction	Layer Name	Features Shape (1D)	Features Shape (2D)	No.Epoch (CNN 1D)	No.Epoch (CNN 2D)
EFFNet	avg_pool	(1, 1280)	(16,16,5)	30	120
	top_conv	(1, 62720)	(16,16,245)	30	120
RESNet152	avg_pool	(1,2048)	(32,32,2)	30	150
NASNetMobile	Prediction	(1,1000)	(2,2,250)	30	60
MobileNetV2	Prediction	(1,1000)	(2,2,250)	30	30

TABLE XIII. FEATURES DIMENSIONS SIZE AND EPOCH DOR SCENE-15 DATASET

Feature Extraction	Layer Name	Features Length (1D)	Features Length (2D)	No.Epoch (CNN 1D)	No.Epoch (CNN 2D)
EFFNET	avg_pool	(1, 1280)	(16,16,5)	120	150
	top_conv	(1, 62720)	(16,16,245)	30	150
RESNET 152	avg_pool	(1,2048)	(32,32,2)	60	150
NASNetMobile	Prediction	(1,1000)	(2,2,250)	60	150
MobileNetV2	Prediction	(1,1000)	(2,2,250)	30	150

Based on Table XIII, the number of epoch required for training the CNN classifiers for Scene-15 dataset was larger than UMS landmark dataset. It was found that the CNN 2D required up to 150 epochs for CNN training.

Fig. 8, Fig. 9, Fig. 10, and Fig. 11 present the graph of model accuracy and model loss over number of epochs for EFFNET and MobilenetV2 by using CNN 2D and CNN 1D classifiers. To determine the appropriate number of epochs for each CNN architecture, the evaluation was made on 30, 60, 90, 120, and 150 epochs. By using 120 number of epoch, the EFFNET with avg\_pool layer managed to obtain the best classification performance with very minimal gap between training and test model lost, as can be seen in Fig. 9. On the

other hand, a slightly larger gap size can be observed between training and testing in model loss in EFFNET using top\_conv layer with stagnant performance in model accuracy despite of larger number of epochs being used as shown in Fig. 11.

In summary, the feature extraction by using EFFNET by using avg\_pool and top\_conv layers with both CNN and SVM classifiers can be considered as the best option in this context and with their own merits. For instance, the EFFNET with avg\_pool layer produced a light feature size which definitely use less computational effort for storage and classification. Meanwhile, the EFFNET with top\_conv layer, even though it produced a larger size of features, but required a very minimum number of epochs to run the CNN classifier with a high classification accuracy. Thus, the trained model, by using EFFNET-avg\_pool with CNN 1D classifier could be deployed in the development of Landmark Recognition System.

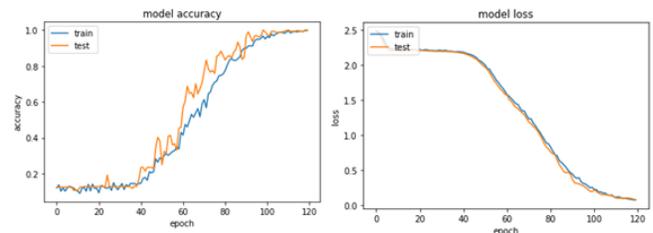


Fig. 8. EFFNET (AVGPOOL)- CNN 2D on UMS scene dataset

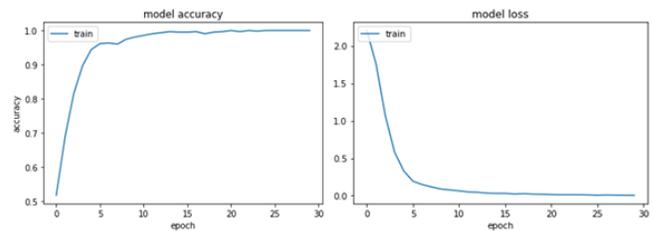


Fig. 9. MobileNetV2 - CNN 1D on UMS scene dataset

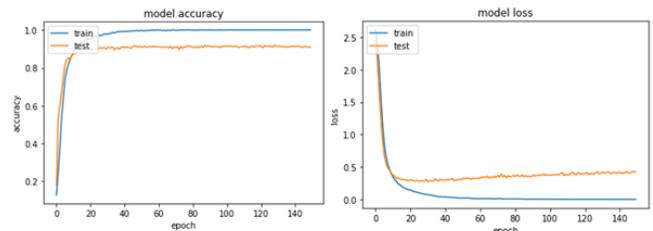


Fig. 10. EFFNET(TOPCONV)- CNN 2D on scene-15 dataset

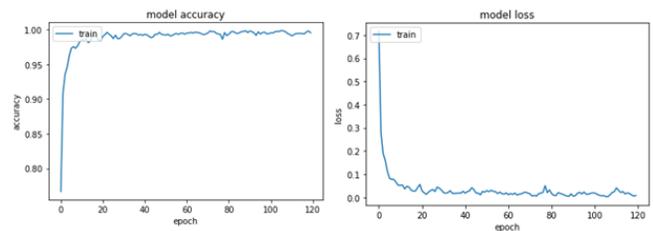


Fig. 11. EFFNET(AVGPOOL)- CNN 1D on scene 15 dataset

C. Effect of Feature Selection

The performance of feature selection methods such as PCA, LDA, Boruta, and RFE on the UMS and Scene-15 datasets, is discussed in this section. The feature selections were applied to EFFNET, RESNET152, NASNETMobile, and MobileNetV2 features, in particular.

1) UMS dataset: The three PCA variations, as shown in Table XIV, mirrored the varying proportions of features selected, as seen in Table IX. As shown in Table XIV, the baseline referred to the findings achieved in the prior trial without any treatment employing feature selection.

TABLE XIV. EFFECT OF PCA ON ACCURACY FOR UMS DATASET

Feature Extraction	Classification	Baseline	PCA 1	PCA 2	PCA 3	Spark-line
EFFNET	LSVM	1.00	1.00	1.00	<b>1.00</b>	
	CNN (1D)	1.00	1.00	1.00	<b>1.00</b>	
	GBDT	1.00	0.99	0.99	0.99	
	SGD	1.00	1.00	1.00	<b>1.00</b>	
	MLP	0.44	0.68	0.59	0.57	
RESNET 152	LSVM	1.00	1.00	1.00	1.00	
	CNN (1D)	1.00	1.00	1.00	1.00	
	GBDT	1.00	0.98	0.96	0.95	
	SGD	0.95	1.00	1.00	1.00	
	MLP	0.12	0.66	0.64	0.64	
NASNet Mobile	LSVM	0.77	0.77	0.77	0.77	
	CNN (1D)	1.00	1.00	1.00	<b>1.00</b>	
	GBDT	0.99	0.91	0.92	0.98	
	SGD	0.82	0.88	0.80	0.91	
	MLP	0.33	0.12	0.12	0.61	
MobileNetV 2	LSVM	0.85	0.85	0.85	0.85	
	CNN (1D)	1.00	1.00	1.00	<b>1.00</b>	
	GBDT	1.00	0.88	0.87	0.89	
	SGD	0.77	0.82	0.77	0.85	
	MLP	0.56	0.72	0.12	0.12	

Based on the overall result in Table XIV, the treatment of PCA had a positive effect on majority of the features as it has retained accuracy performance and even more, slight improvement on the accuracy can be observed on all the features especially the MLP classifiers. In a flipside, the accuracy performance using GBDT has slightly affected regardless any features used.

Table XV shows the classification performance after the LDA and Boruta were performed on all the features. The LDA had a positive effect on the accuracy performance for almost all the features except the classification using MLP. Despite pruning more than 90% of features by using LDA, the accuracy performance improvement can be observed on RESNET152, NASNETMobile and MobileNetV2 along sustaining the best accuracy performance on EFFNET. On the other hand, the BORUTA only demonstrated positive effect on EFFNET and RESNET152. The other highlight was the classification using MLP on EFFNET features has dramatically improved the accuracy performance from 0.44 to 0.83.

TABLE XV. EFFECTS OF LDA AND BORUTA ON ACCURACY FOR UMS DATASET

Feature Extraction	Classification	Baseline	LD A	BORUT A	Trendline
EFFNET	LSVM	1.00	<b>1.00</b>	1.00	
	CNN (1D)	1.00	<b>1.00</b>	1.00	
	GBDT	1.00	<b>1.00</b>	1.00	
	SGD	1.00	<b>1.00</b>	1.00	
	MLP	0.44	0.19	0.83	
RESNET 152	LSVM	1.00	<b>1.00</b>	1.00	
	CNN (1D)	1.00	<b>1.00</b>	1.00	
	GBDT	1.00	<b>1.00</b>	1.00	
	SGD	0.95	<b>1.00</b>	0.99	
	MLP	0.12	0.12	0.12	
NASNet Mobile	LSVM	0.77	<b>1.00</b>	0.77	
	CNN (1D)	1.00	<b>1.00</b>	1.00	
	GBDT	0.99	<b>1.00</b>	0.99	
	SGD	0.82	<b>1.00</b>	0.93	
	MLP	0.33	0.12	0.12	
MobileNetV 2	LSVM	0.85	<b>1.00</b>	0.85	
	CNN (1D)	1.00	<b>1.00</b>	1.00	
	GBDT	1.00	<b>1.00</b>	1.00	
	SGD	0.77	<b>1.00</b>	0.77	
	MLP	0.56	0.19	0.12	

Table XVI presents the analysis of feature selection performance using RFE.

TABLE XVI. EFFECTS OF RFE ON ACCURACY FOR UMS DATASET

Feature Extraction	Classification	Baseline	RFE 1	RFE 2	RFE 3	Spark-line
EFFNet	LSVM	1.00	1.00	1.00	<b>1.00</b>	
	CNN (1D)	1.00	1.00	1.00	<b>1.00</b>	
	GBDT	1.00	1.00	1.00	<b>1.00</b>	
	SGD	1.00	1.00	1.00	<b>1.00</b>	
	MLP	0.44	0.46	0.48	0.55	
RESNet152	LSVM	1.00	1.00	1.00	1.00	
	CNN (1D)	1.00	1.00	1.00	1.00	
	GBDT	1.00	1.00	1.00	0.96	
	SGD	0.95	0.93	1.00	1.00	
	MLP	0.12	0.39	0.12	0.64	
NASNet Mobile	LSVM	0.77	0.14	0.16	0.20	
	CNN (1D)	1.00	0.82	0.34	0.12	
	GBDT	0.99	0.96	0.98	0.99	
	SGD	0.82	0.44	0.39	0.10	
	MLP	0.33	0.12	0.12	0.12	
MobileNet V2	LSVM	0.85	0.13	0.23	0.22	
	CNN (1D)	1.00	1.00	0.77	0.17	
	GBDT	1.00	0.99	0.99	0.98	
	SGD	0.77	0.13	0.13	0.12	
	MLP	0.56	0.12	0.12	0.12	

RFE worked well on EFFNET and RESNET152, as shown in Table XVII. Moreover, the performance of MLP on RESNET152 had substantially improved from 0.12 to 0.64. On the other hand, RFE absolutely failed to perform on NASNetMobile and MobileNetV2, resulting in a significant fall in the accuracy of all classifiers used.

Next, the detailed analysis of feature selection performance on each feature and machine learning classifier are shown in Fig. 12, Fig. 13, Fig.14 and Fig.15.

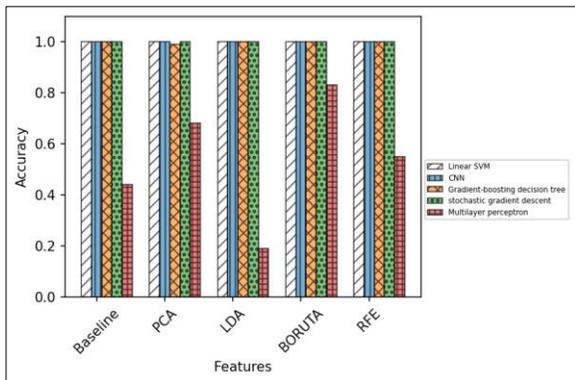


Fig. 12. Effects of feature selection on machine learning classifiers for EFFNET (UMS dataset)

As shown in Fig. 12, except for MLP, all classifiers in EFFNET performed remarkably well on all feature selections. Whereby, the EFFNET features would be more compatible with MLP if PCA and BORUTA is applied as the accuracies were increased by 55% and 89% respectively. On RESNET152 with EFFNET, a similar pattern of feature selection performance can be observed, as shown in Fig. 13. In fact, regardless of which feature selection is employed, the accuracy of SGD can be improved. When PCA and RFE were used with MLP, a positive effect on accuracy was noticed.

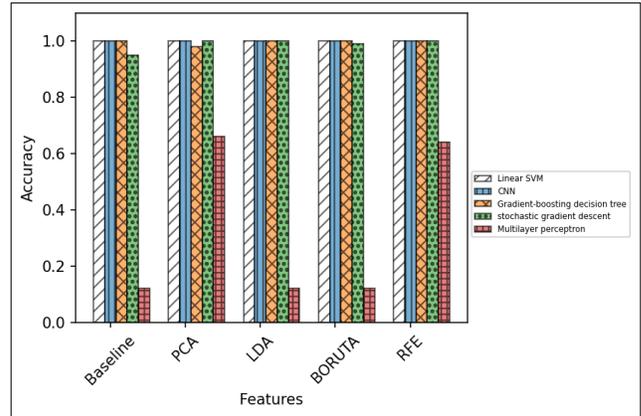


Fig. 13. Effects of feature selection on machine learning classifiers for RESNET152 (UMS dataset)

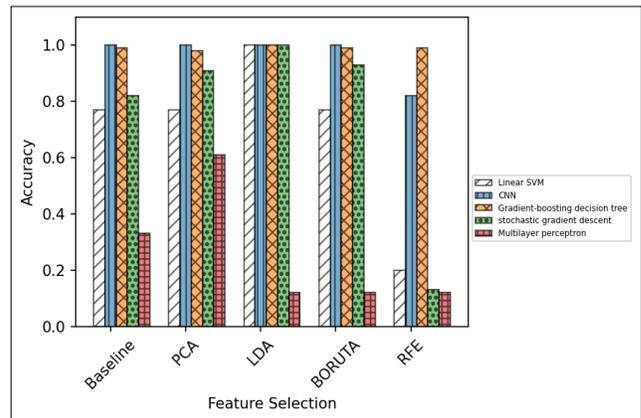


Fig. 14. Effects of feature selection on machine learning classifiers for NASNETMOBILE (UMS dataset)

According to the graph in Fig. 14, GBDT's performance appeared to be consistent across all feature selections, but the performance of the other classifiers dropped when RFE was applied. The best performance of LSVM and SGD could be seen when LDA was used. On MobileNetV2, CNN performed very well with all the feature selections and GBDT was slightly incompatible with PCA. Similar with NASNetMobile, LDA had also improved the accuracy of LSVM and SGD.

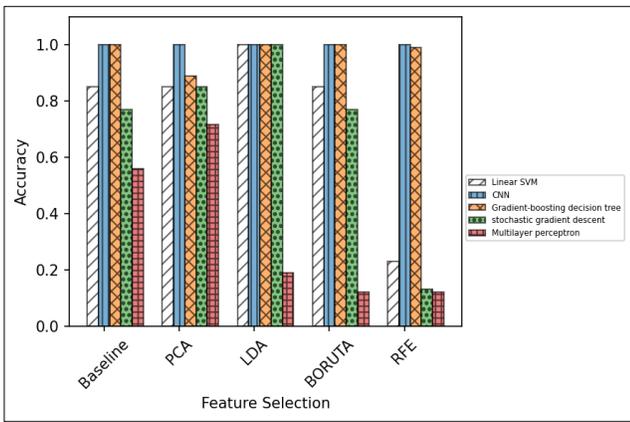


Fig. 15. Effects of feature selection on machine learning classifiers for MOBILENETV2 (UMS dataset)

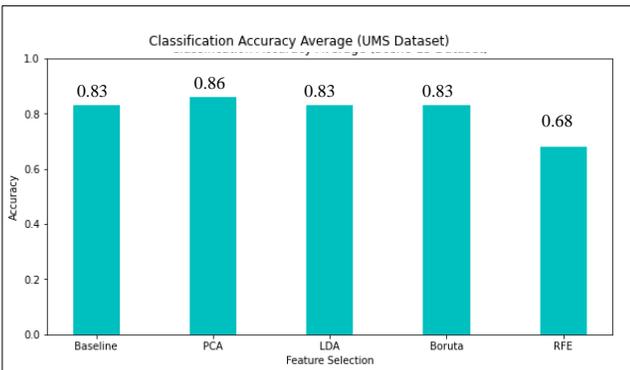


Fig. 16. The average performance comparisons of feature selection for UMS dataset

The summary of feature selection performance across features and classifiers for the UMS dataset is shown in Fig. 16. The PCA was found to be the most robust feature selection method since its performance was consistent across various features and classifiers. However, when accuracy and feature size were taken into account, LDA's performance was the most significant. Meanwhile, if execution time was not a major concern and automatic feature selection is one of the criteria for selecting features, the BORUTA could be considered. Aside from that, the results of Tables XII, XIII, and XIV implied that EFFNET is the best and stable features. The best classifiers were GBDT and CNN, which consistently excelled across a variety of feature selections.

2) *Scene-15 dataset*: Table XVII shows the performance analysis of PCA on Scene-15 dataset.

TABLE XVII. EFFECTS OF PCA ON ACCURACY FOR SCENE-15 DATASET

Feature Extraction	Classification	Baseline	PCA 1	PCA 2	PCA 3	Spark line
EFFNET	LSVM	0.94	0.94	0.93	0.91	
	CNN (1D)	0.94	0.93	0.92	0.93	
	GBDT	0.68	0.06	0.01	0.05	
	SGD	0.68	0.93	0.93	0.93	
	MLP	0.43	0.56	0.32	0.33	
RESNET15 2	LSVM	0.62	0.61	0.59	0.55	
	CNN (1D)	0.62	0.66	0.66	0.66	
	GBDT	0.41	0.33	0.44	0.49	
	SGD	0.37	0.54	0.52	0.52	
	MLP	0.23	0.40	0.44	0.44	
NASNet Mobile	LSVM	0.68	0.70	0.68	0.67	
	CNN (1D)	0.74	0.77	0.73	0.71	
	GBDT	0.55	0.52	0.51	0.54	
	SGD	0.58	0.68	0.61	0.64	
	MLP	0.39	0.63	0.08	0.08	
MobileNet V2	LSVM	0.69	0.70	0.70	0.68	
	CNN (1D)	0.82	0.79	0.79	0.73	
	GBDT	0.07	0.42	0.42	0.56	
	SGD	0.68	0.65	0.68	0.65	
	MLP	0.34	0.60	0.08	0.08	
<b>AVERAGE</b>		<b>0.57</b>	<b>0.62</b>	<b>0.55</b>	<b>0.57</b>	

Overall, PCA did not enhance classification accuracy considerably. SGD and MLP are the only two classifiers that performed better with PCA. For instance, EFFNET-SGD accuracy increased from 0.68 to 0.94, whereas NASNETMobile's classification accuracy increased from 0.39 to 0.63.

The accuracy performance of LDA and BORUTA treatment as compared to without feature selection treatment (Baseline) can be referred in Table XVIII. As depicted in Table XVIII, LDA performed excellently on many features and classifiers, except EFFNET-GBDT, NASNETMobile-GBDT and MOBILENetV2-GBDT. In contrast, BORUTA did not increase the accuracy of nearly all features, and there was even a slight drop in accuracy.

TABLE XVIII. EFFECTS OF LDA AND BORUTA ON ACCURACY FOR SCENE-15 DATASET

Feature Extraction	Classification	Baseline	LDA	BORUTA	Spark line
EFFNet	LSVM	0.94	0.99	0.93	
	CNN (1D)	0.94	0.99	0.93	
	GBDT	0.68	0.05	0.70	
	SGD	0.68	0.99	0.90	
	MLP	0.43	0.69	0.64	
RESNet152	LSVM	0.62	0.93	0.58	
	CNN (1D)	0.62	0.93	0.60	
	GBDT	0.41	0.69	0.38	
	SGD	0.37	0.94	0.36	
	MLP	0.23	0.58	0.08	
NASNet Mobile	LSVM	0.68	0.91	0.40	
	CNN (1D)	0.74	0.90	0.72	
	GBDT	0.55	0.04	0.55	
	SGD	0.58	0.89	0.42	
	MLP	0.39	0.77	0.46	
MobileNetV2	LSVM	0.69	0.75	0.62	
	CNN (1D)	0.82	0.91	0.79	
	GBDT	0.07	0.01	0.19	
	SGD	0.68	0.91	0.61	
	MLP	0.34	0.75	0.23	
<b>AVERAGE</b>		<b>0.57</b>	<b>0.73</b>	<b>0.55</b>	

The analysis of RFE accuracy performance is shown in Table XIX. The pattern of data presented in Table XIX obviously indicates that RFE has brought less impact on improving almost all feature representation. However, the positive effects of RFE can be seen on EFFNET-SGD, EFFNET-MLP and MobileNetV2-MLP.

TABLE XIX. EFFECTS OF RFE ON ACCURACY FOR SCENE-15 DATASET

Feature Extraction	Classification	Baseline	RFE 1	RFE 2	RFE 3	Spark line
EFFNet	LSVM	0.94	0.94	0.93	0.89	
	CNN (1D)	0.94	0.92	0.93	0.89	
	GBDT	0.68	0.61	0.70	0.66	
	SGD	0.68	0.91	0.92	0.84	
	MLP	0.43	0.54	0.36	0.38	
RESNet152	LSVM	0.62	0.59	0.46	0.04	
	CNN (1D)	0.62	0.57	0.53	0.11	
	GBDT	0.41	0.40	0.37	0.12	
	SGD	0.37	0.34	0.45	0.08	
	MLP	0.23	0.58	0.08	0.08	

	MLP	0.23	0.24	0.08	0.08	
NASNet Mobile	LSVM	0.68	0.19	0.04	0.04	
	CNN (1D)	0.74	0.79	0.63	0.20	
	GBDT	0.55	0.66	0.56	0.07	
	SGD	0.58	0.44	0.08	0.06	
	MLP	0.39	0.27	0.08	0.08	
MobileNet V2	LSVM	0.69	0.70	0.68	0.06	
	CNN (1D)	0.82	0.80	0.80	0.60	
	GBDT	0.07	0.07	0.07	0.19	
	SGD	0.68	0.67	0.69	0.23	
	MLP	0.34	0.59	0.08	0.08	
<b>AVERAGE</b>		<b>0.57</b>	<b>0.56</b>	<b>0.47</b>	<b>0.28</b>	

Fig. 17 to 20 show a detailed analysis of feature selection performance for each feature and machine learning classifier. Based on the graph shown in Fig. 17, the transformation of EFFNET feature by using LDA had improved the classification accuracy of LSVM, CNN, SGD and MLP. In addition to that, the PCA, BORUTA and RFE brought significant effects on the accuracies for MLP and SGD.

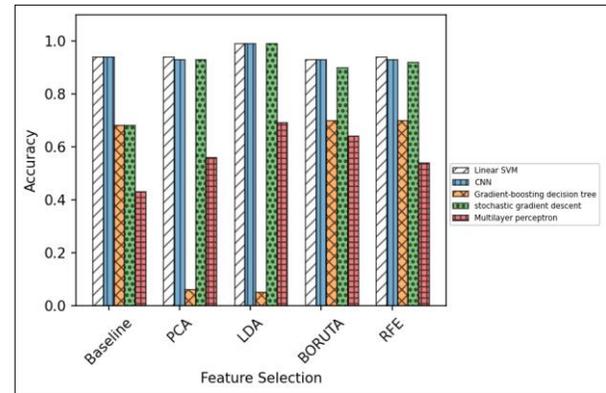


Fig. 17. Effects of feature selection on machine learning classifiers for EFFNET (scene-15 dataset)

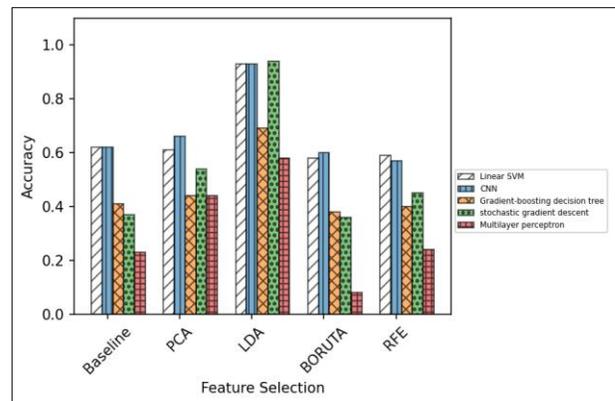


Fig. 18. Effects of feature selection on machine learning classifiers for RESNET152 (scene-15 dataset)

As for RESNET152, as shown in Fig. 18, there was a tremendous increase on the accuracy when LDA was being used to transform the features for CNN, LSVM and SGD. The rest of the feature selection techniques by using PCA, BORUTA and RFE seemed to have less positive impacts on the accuracies. Similarly in Fig. 19 and Fig. 20, LDA still outperformed the accuracy of PCA, BORUTA and RFE on all classifiers except GBDT. For NASNETMobile, PCA demonstrated a bit of an improvement on the accuracies for CNN, SGD and MLP. There were no positive effects on the LSVM, CNN, SGD, and MLP accuracies for BORUTA and RFE.

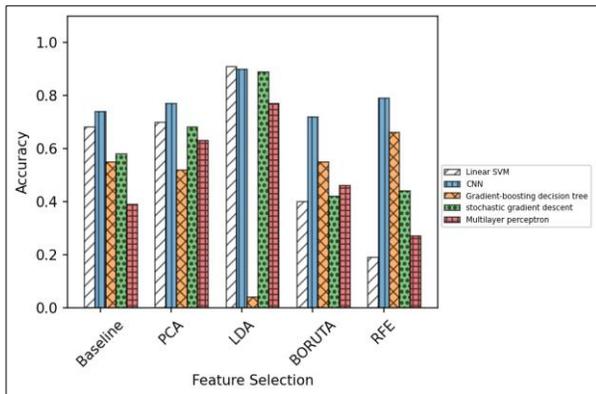


Fig. 19. Effect of feature selection on machine learning classifiers for NASNETMOBILE (scene-15 dataset)

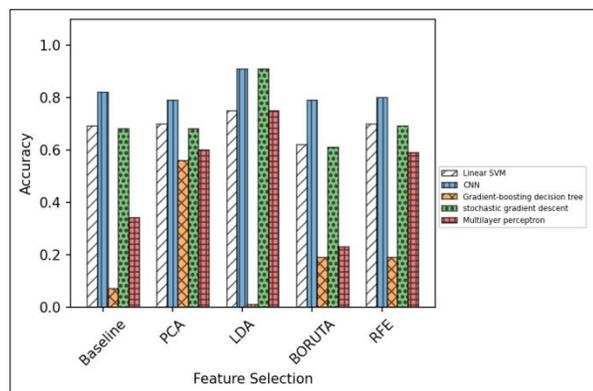


Fig. 20. Effect of feature selection on machine learning classifiers for MOBILENETV2 (scene-15 dataset)

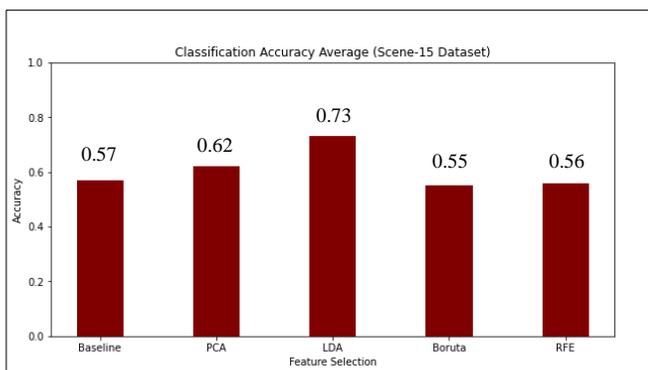


Fig. 21. The average performance comparisons of feature selection for scene-15 dataset

Fig. 21 shows the summary of feature selection performance on the Scene-15 dataset. LDA was the best feature selection technique for the Scene-15 dataset since it not only worked with a wide range of features and classifiers, but it also improved classification accuracy significantly. BORUTA and RFE, on the other hand, have no substantial impact on classification performance. Due to the constant performance across numerous feature selections, it can also be inferred that EFFNET is the best features and, LSVM is the best classifier.

## V. CONCLUSION AND FUTURE WORKS

This paper evaluated several transfer learning approaches and feature selections for effective and super lightweight landmark recognition model. A landmark recognition model was trained through the features extraction by using the pre-trained CNN architectures and machine learning classifiers. The new UMS landmark datasets were created, and the landmark recognition model was also evaluated with the Scene-15 dataset. The findings showed that the EFFNET CNN architecture with CNN classifier is the best feature extraction and classifier in this study. EFFNET-CNN achieved 100% and 94.26% accuracies on UMS landmark and Scene-15 dataset, respectively. Moreover, the features created by EFFNET were more compact compared to the other features. Furthermore, based on the evaluation of several feature selection algorithms, LDA was determined to be the best feature selection technique for vastly reducing feature dimensionality by 99.69% for UMS landmark dataset and 98.90% for Scene-15 dataset while maintaining good accuracies. However, although a super lightweight landmark recognition model was produced, it must undergo extra pre-processing step to reduce the dimensionality of features which will impose excessive computational costs of processing. Therefore, future works that can be suggested are to evaluate the effect of the proposed dimensionality reduction technique on the computational cost of the algorithms as well as to test it on various benchmark datasets.

## REFERENCES

- [1] L. Xie, F. Lee, L. Liu, K. Kotani, and Q. Chen, "Scene recognition: A comprehensive survey," *Pattern Recognit.*, vol. 102, 2020, doi: 10.1016/j.patcog.2020.107205.
- [2] A. Boiarov and E. Tyantov, "Large scale landmark recognition via deep metric learning," *Int. Conf. Knowl. Manag. Proc.*, pp. 169–178, 2019, doi: 10.1145/3357384.3357956.
- [3] M. Jiafa, W. Weifeng, H. Yahong, and S. Weiguo, "A scene recognition algorithm based on deep residual network," *Syst. Sci. Control Eng.*, vol. 7, no. 1, pp. 243–251, 2019, doi: 10.1080/21642583.2019.1647576.
- [4] N. M. Firdaus, D. Chahyati, and M. I. Fanany, "Tourist attractions classification using ResNet," *2018 Int. Conf. Adv. Comput. Sci. Inf. Syst. ICACSIS 2018*, pp. 429–433, 2019, doi: 10.1109/ICACSIS.2018.8618235.
- [5] M. Jiafa, W. Weifeng, H. Yahong, and S. Weiguo, "A scene recognition algorithm based on deep residual network," *Syst. Sci. Control Eng.*, vol. 7, no. 1, pp. 243–251, 2019, doi: 10.1080/21642583.2019.1647576.
- [6] M. N. Razali, A. S. Shafie, and R. Hanapi, "Performance Evaluation of Masked Face Recognition Using Deep Learning for Covid-19 Standard of Procedure (SOP) Compliance Monitoring," *2021 6th IEEE Int. Conf. Recent Adv. Innov. Eng. ICRAIE 2021*, vol. 2021, 2021, doi: 10.1109/ICRAIE52900.2021.9703986.
- [7] M. N. Razali, A. S. Shafie, and R. Hanapi, "Performance Evaluation of Masked Face Recognition Using Deep Learning for Covid-19 Standard of Procedure (SOP) Compliance Monitoring," *vol. 2021*, pp. 1–7, 2022, doi: 10.1109/icraie52900.2021.9703986.

- [8] [8] S. Khan et al., "DeepSmoke: Deep learning model for smoke detection and segmentation in outdoor environments," *Expert Syst. Appl.*, vol. 182, no. December 2020, p. 115125, 2021, doi: 10.1016/j.eswa.2021.115125.
- [9] E. Uçar, Ü. Atila, M. Uçar, and K. Akyol, "Automated detection of Covid-19 disease using deep fused features from chest radiography images," *Biomed. Signal Process. Control*, vol. 69, no. December 2020, p. 102862, 2021, doi: 10.1016/j.bspc.2021.102862.
- [10] N. Jahan, M. S. Anower, and R. Hassan, "Automated Diagnosis of Pneumonia from Classification of Chest X-Ray images using EfficientNet," 2021 Int. Conf. Inf. Commun. Technol. Sustain. Dev. ICICT4SD 2021 - Proc., pp. 235–239, 2021, doi: 10.1109/ICICT4SD50815.2021.9397055.
- [11] M. M. A. Monshi, J. Poon, V. Chung, and F. M. Monshi, "CovidXrayNet: Optimizing data augmentation and CNN hyperparameters for improved COVID-19 detection from CXR," *Comput. Biol. Med.*, vol. 133, no. March, p. 104375, 2021, doi: 10.1016/j.compbiomed.2021.104375.
- [12] A. A. Pokroy and A. D. Egorov, "EfficientNets for DeepFake Detection: Comparison of Pretrained Models," *Proc. 2021 IEEE Conf. Russ. Young Res. Electr. Electron. Eng. ElConRus 2021*, pp. 598–600, 2021, doi: 10.1109/ElConRus51938.2021.9396092.
- [13] A. S. Timmaraju and A. Chatterjee, "Monulens : Real-time mobile-based Landmark Recognition."
- [14] Meiliana, D. Irmanti, M. R. Hidayat, N. V. Amalina, and D. Suryani, "Mobile Smart Travelling Application for Indonesia Tourism," *Procedia Comput. Sci.*, vol. 116, pp. 556–563, 2017, doi: 10.1016/j.procs.2017.10.059.
- [15] A. Crudge, W. Thomas, and K. Zhu, "Landmark Recognition Using Machine Learning," pp. 1–5, 2014.
- [16] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, "Learning deep features for scene recognition using places database," *Adv. Neural Inf. Process. Syst.*, vol. 1, no. January, pp. 487–495, 2014.
- [17] N. Shigei, K. Mandai, S. Sugimoto, R. Takaesu, and Y. Ishizuka, "Land-use classification using convolutional neural network with bagging and reduced categories," *Lect. Notes Eng. Comput. Sci.*, vol. 2239, pp. 7–11, 2019.
- [18] E. G. Mounq, C. J. Hou, M. M. Sufian, M. H. A. Hijazi, J. A. Dargham, and S. Omatu, "Fusion of moment invariant method and deep learning algorithm for COVID-19 classification," *Big Data Cogn. Comput.*, vol. 5, no. 4, 2021, doi: 10.3390/bdcc5040074.
- [19] V. Parikh, M. Keskar, D. Dharia, and P. Gotmare, "A Tourist Place Recommendation and Recognition System," *Proc. Int. Conf. Inven. Commun. Comput. Technol. ICICCT 2018*, no. Iccict, pp. 218–222, 2018, doi: 10.1109/ICICCT.2018.8473077.
- [20] UMS, "Landmarks in UMS," 2021. <https://www.ums.edu.my/v5/en/landmark-of-ums>.
- [21] E. O. Nixon and M. N. Razali, "Ums Landmark Recognition Dataset," 2022. <https://doi.org/10.34740/KAGGLE/DS/1877538>.
- [22] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2, pp. 2169–2178, 2006, doi: 10.1109/CVPR.2006.68.
- [23] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," *36th Int. Conf. Mach. Learn. ICML 2019*, vol. 2019-June, pp. 10691–10700, 2019.
- [24] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2016-Decem, pp. 770–778, 2016, doi: 10.1109/CVPR.2016.90.
- [25] B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le, "Learning Transferable Architectures for Scalable Image Recognition," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pp. 8697–8710, 2018, doi: 10.1109/CVPR.2018.00907.
- [26] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. C. Chen, "MobileNetV2: Inverted Residuals and Linear Bottlenecks," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pp. 4510–4520, 2018, doi: 10.1109/CVPR.2018.00474.
- [27] M. M. A. Monshi, J. Poon, V. Chung, and F. M. Monshi, "CovidXrayNet: Optimizing data augmentation and CNN hyperparameters for improved COVID-19 detection from CXR," *Comput. Biol. Med.*, vol. 133, no. December 2020, p. 104375, 2021, doi: 10.1016/j.compbiomed.2021.104375.
- [28] M. N. Razali et al., "Indigenous food recognition model based on various convolutional neural network architectures for gastronomic tourism business analytics," *Inf.*, vol. 12, no. 8, 2021, doi: 10.3390/info12080322.
- [29] V. K. Shrivastava, M. K. Pradhan, and M. P. Thakur, "Neural Networks for Rice Plant Disease Classification," pp. 1023–1030, 2021.
- [30] C. Cordoş, L. Mihailă, P. Faragó, and S. Hintea, "ECG signal classification using Convolutional Neural Networks for Biometric Identification," pp. 167–170, 2021.
- [31] R. V. M. Da Nóbrega, S. A. Peixoto, S. P. P. Da Silva, and P. P. R. Filho, "Lung Nodule Classification via Deep Transfer Learning in CT Lung Images," *Proc. - IEEE Symp. Comput. Med. Syst.*, vol. 2018-June, pp. 244–249, 2018, doi: 10.1109/CBMS.2018.00050.
- [32] S. A. A. Ahmed, B. Yanikoglu, O. Goksu, and E. Aptoula, "Skin Lesion Classification with Deep CNN Ensembles," *2020 28th Signal Process. Commun. Appl. Conf. SIU 2020 - Proc.*, pp. 1–4, 2020, doi: 10.1109/SIU49456.2020.9302125.
- [33] N. Merrin Prasanna, D. Subash Chandra Mouli, G. Sireesha, K. Priyanka, D. Radha, and B. Manmadha, "Classification of food categories and ingredients approximation using an fd-mobilenet and TF-Yolo," *Int. J. Adv. Sci. Technol.*, vol. 29, no. 6, pp. 3104–3114, 2020.
- [34] M. N. S. Zainudin, N. Sulaiman, N. Mustapha, T. Perumal, and R. Mohamed, "Two-stage feature selection using ranking self-adaptive differential evolution algorithm for recognition of acceleration activity," *Turkish J. Electr. Eng. Comput. Sci.*, vol. 26, no. 3, pp. 1378–1389, 2018, doi: 10.3906/elk-1709-138.
- [35] L. P. Hung, R. Alfred, and M. H. A. Hijazi, "Comparison of feature selection methods for sentiment analysis," *Commun. Comput. Inf. Sci.*, vol. 872, no. February 2020, pp. 261–272, 2018, doi: 10.1007/978-3-319-96292-4\_21.
- [36] M. N. Razali, N. Manshor, A. A. Halin, N. Mustapha, and R. Yaakob, "Extremal Region Selection for MSER Detection in Food Recognition," *ASM Sci. J.*, vol. 15, no. May, pp. 1–11, 2021, doi: 10.32802/ASMSCJ.2020.485.
- [37] D. J. Bartholomew, "Principal components analysis," *Int. Encycl. Educ.*, pp. 374–377, 2010, doi: 10.1016/B978-0-08-044894-7.01358-0.
- [38] F. Song, D. Mei, and H. Li, "Feature selection based on linear discriminant analysis," *Proc. - 2010 Int. Conf. Intell. Syst. Des. Eng. Appl. ISDEA 2010*, vol. 1, pp. 746–749, 2010, doi: 10.1109/ISDEA.2010.311.
- [39] M. B. Kursu, A. Jankowski, and W. R. Rudnicki, "Boruta - A system for feature selection," *Fundam. Informaticae*, vol. 101, no. 4, pp. 271–285, 2010, doi: 10.3233/FI-2010-288.
- [40] X. Zeng, Y. W. Chen, C. Tao, and D. Van Alphen, "Feature selection using recursive feature elimination for handwritten digit recognition," *IIH-MSP 2009 - 2009 5th Int. Conf. Intell. Inf. Hiding Multimed. Signal Process.*, pp. 1205–1208, 2009, doi: 10.1109/IIH-MSP.2009.145.

# Performance Comparison of the Kernels of Support Vector Machine Algorithm for Diabetes Mellitus Classification

Dimas Aryo Anggoro, Dian Permatasari

Informatics Department, Universitas Muhammadiyah Surakarta  
Surakarta, Indonesia

**Abstract**—Diabetes Mellitus is a disease where the body cannot use insulin properly, so this disease is one of the health problems in various countries. Diabetes Mellitus can be fatal, cause other diseases, and even lead to death. Based on this, it is essential to have prediction activities to find out a disease. The SVM algorithm is used in classifying Diabetes Mellitus diseases. This study aimed to compare the accuracy, precision, recall, and F1-Score values of the SVM algorithm with various kernels and data preprocessing. Data preprocessing included data splitting, normalization, and data oversampling. This research has the benefit of solving health problems based on the percentage of Diabetes Mellitus and can be used as material for accurate information. The results of this study are that the highest accuracy was obtained by 80% (obtained from the polynomial kernel), the highest precision was obtained by 65%, which was also obtained from the polynomial kernel, and the highest recall was obtained by 79% (obtained from the RBF kernel) and the highest F1-score was obtained by 70% (which was also obtained from the RBF kernel).

**Keywords**—Diabetes mellitus; kernel; normalization; oversampling; SVM

## I. INTRODUCTION

Diabetes Mellitus is a disease where blood sugar levels are overly high because the body cannot use insulin properly. Diabetes Mellitus has become a severe health problem in various countries, including Indonesia [1]. The International Diabetes Federation (IDF) explained that in 2021 the number of people with Diabetes Mellitus in Indonesia reached 19.5 million, while in 2019 the figure was 10.7 million. This means there has been an increase of nearly 9 million cases in just two years, during the COVID-19 pandemic. With almost two times the addition, Indonesia ranks fifth in the world. Not only in Indonesia, but this upward trend in cases also occurs worldwide. According to IDF data, at least 1 in 10 people or as many as 537 million people live with Diabetes Mellitus. If not appropriately treated immediately, Diabetes Mellitus can be fatal, cause other diseases, and even lead to death. Based on this, it is essential to have prediction activities to find out a disease. This activity is carried out so a disease can be detected quickly and treated immediately.

Activities in predicting various diseases have been carried out in various scientific fields, one of which is computer science. Along with the development of information and communication technology, it can be used to improve the

ability of the system to help detect Diabetes Mellitus disease [2]. Data mining is part of the Knowledge Discovery in Database (KDD) process that can classify, predict, and get a lot of information from large data sets [3]. Classification is an important stage in data mining, classification is carried out by looking at variables from existing data groups and aims to predict the class of an object that was not previously known [4].

## II. LITERATURE REVIEW

Previous research regarding applying the K-Nearest Neighbour classification model to the diabetes patient dataset explained that the study had the highest accuracy of 39% [5]. Another study is the implementation of the Decision Tree C4.5 algorithm for diabetes prediction resulted in a prediction model with the highest accuracy of 70.32% [6]. The previous study's shortcoming is that the prediction model's accuracy is still below 80%, so there is a need to improve accuracy performance. In research [7] that compared the accuracy, recall, and precision classification of the C4.5 algorithm, Random Forest, Support Vector Machine (SVM), and Naïve Bayes resulted in the C4.5 algorithm obtaining accuracy of 86.67%, the Random Forest algorithm obtained an accuracy of 83.33%, the SVM algorithm obtained accuracy by 95%, and the Naive Bayes algorithm obtained an accuracy of 86.67%. The highest accuracy algorithm is the SVM algorithm. Therefore in this study applying the SVM algorithm for the classification of Diabetes Mellitus disease.

The SVM algorithm was chosen because it is reliable in processing large amounts of data by optimizing hyperplanes in high-dimensional space that maximizes margins between data [8]. The kernel in SVM is used to determine kernel parameters and produce the best accuracy in the classification process. Linear kernels are used when a hyperplane can easily separate classified data. At the same time, non-linear kernels are used when the data is separated using curved lines or a plane in space with high dimensions [9].

This study aims to compare the performance metrics e.g., accuracy, precision, recall, and F1-Score values of the SVM algorithm with various kernels and preprocessing data in the classification of Diabetes Mellitus disease. The SVM algorithm is evaluated to determine which kernel can produce the best performance metric.

It has the benefit of solving health problems based on the percentage of Diabetes Mellitus and can be an accurate information material. The output of this study is to imply that the SVM algorithm is expected to show better performance values than previous studies.

### III. METHODOLOGY

#### A. Data Collection

The first stage in this study is the collection of Diabetes Mellitus datasets. The dataset used is the Pima Indian diabetes dataset obtained from the UCI Machine Learning Repository. Several variables and attributes can facilitate the research process in data mining. The Pima Indian diabetes dataset consists of 768 data and 9 features. The variables and features used are shown in Table I.

TABLE I. VARIABLES AND ATTRIBUTES OF PEOPLE WITH DIABETES MELLITUS

Variable	Attribute
X1	Pregnancies, the number of pregnancies during life in the range of 0-17 times.
X2	Glucose, glucose/blood sugar levels. Normal blood sugar levels are below 120 mg/dL, while the sugar levels of diabetics are more than 120 mg/dL. The data range in the dataset is 0-199 mg/dL.
X3	Blood Pressure, blood pressure with mmHg units, the data range in the dataset is 0-112 mmHg.
X4	Skin Thickness, skin fold thickness with a data range of 0-99 mm. The norm is about 12.5 mm.
X5	Insulin, insulin levels in the blood with a data range of 0-846 U / ml.
X6	BMI, body mass weight with a data range of 0-67.1 BMI
X7	Diabetes Pedigree Function, History of diabetes Mellitus disease in the family with a data range of 1.001-2.42.
X8	Age, age of the patient (years) with a data range of 21-81 years.
Y	Outcome, negative and positive class variables (0 and 1). 0 are indicators of non-diabetics while 1 is an indicator of diagnosed diabetics.

#### B. Data Preprocessing

1) *Data splitting*: The next stage is the data splitting stage, which separates training data and testing data. Training data is used to create models that are applied to testing data [10] and testing data cannot be used for the training process, so the model learns from the new data [11]. Training and testing data are determined randomly, so the proportion between categories remains balanced [12]. In this study, splitting data was divided into 80% training data and 20% testing data.

2) *Data normalization*: Normalization of data in datasets aims to create data in the same range of values [13]. This study used the min-max and z-score normalization methods.

a) *Min-Max normalization*: Normalization of min-max can overcome non-uniform data forms with a range of values greater than 0-1 [14]. Min-max normalization was chosen because it has the advantage that the data is balanced between before and after normalization [15]. The normalization of min-max is presented in (1).

$$x_{new} = \frac{x_{old} - x_{min}}{x_{max} - x_{min}} \quad (1)$$

$x_{new}$  represents the min-max value,  $x_{old}$  is the value to be normalized,  $x_{min}$  is the lowest value of the overall data and  $x_{max}$  is the highest value of the entire data.

b) *Z-Score normalization*: Z-Score normalization is used to compare the performance or quality of data goals with the average distribution of data across groups based on standard deviation values [16]. Z-score normalization was chosen because it is a suitable method for balancing the data scale [17]. (2) is a formula for knowing the z-score.

$$x_{new} = \frac{x_{old} - \mu}{\sigma} \quad (2)$$

$x_{new}$  is the z-score value,  $x_{old}$  is the value to be normalized,  $\mu$  is the average value of the whole data and  $\sigma$  is the standard deviation value.

3) *SMOTE (Synthetic Minority Over-sampling Technique)*: The SMOTE method can handle dataset class imbalances by making data replication of minor classes equivalent to major classes [18]. The diabetes dataset used in this study had a total of 268 positive classes and 500 negative classes, so there was an imbalance between the positive and negative classes. Therefore, the SMOTE method was used in this study to balance between positive classes and negative classes. (3) is the formula for SMOTE.

$$x_{syn} = x_i + (x_{knn} - x_i)\gamma \quad (3)$$

$x_{syn}$  is the resulting new class data,  $x_i$  is the approach to i,  $x_{knn}$  is the x closest to  $x_i$  and  $\gamma$  is a random number between 0-1.

#### C. Data Processing

1) *Support Vector Machine (SVM)*: SVM is a good algorithm for data classification [19] with the principle of finding the best hyperplane that serves as a separator of two data classes [20]. The best hyperplane is determined by measuring the hyperplane margin and finding its maximum point, margin is the distance between the hyperplane and the nearest point of each class and this closest point is called the support vector [21]. The following is a description of SVM, there is data  $\vec{x}i \in (\vec{x}1, \vec{x}2 \vec{x}3... \vec{x}n)$   $x_i$  is data consisting of n attributes and two classes  $y_i \in +1, -1$ . Suppose that the two classes can be perfectly separated by a d-dimensional hyperplane defined by (4).

$$\vec{w} \cdot \vec{x}i + b = 0 \quad (4)$$

Data  $\vec{x}i$  which belonging to the positive class (+1) are shown in (5).

$$\vec{w} \cdot \vec{x}i + b \geq -1 \quad (5)$$

Meanwhile, data  $\vec{x}i$  belonging to the negative class (-1) are shown in (6).

$$\vec{w} \cdot \vec{x}i + b \leq +1 \quad (6)$$

The maximum margin can be obtained by maximizing the value of the distance between the hyperplane and its closest point or support vector which  $\frac{1}{\|\vec{w}\|}$  [22]. It is formulated as Quadratic Programming (QP) by looking for a minimum point based on (7).

$$\min \tau(w) = \frac{1}{2} \|\vec{w}\|^2 \quad (7)$$

By paying attention to the constraints on (8).

$$y_i(\vec{x}_i, \vec{w} + b) - 1 \geq 0 \quad (8)$$

$y_i$  is the target class to  $i$ ,  $\vec{x}_i$  is the input data to  $i$ ,  $\vec{w}$  is the weight, and  $b$  is the relative field position.

2) *Kernel SVM*: To work around high-dimensional data, a kernel can transform the input space into a feature space [23]. Kernel functions commonly used in SVM are Linear [24], Radial Basic Function (RBF) and Polynomial [25]. The parameters possessed by kernel functions are used in the testing process [26]. There is no definite conclusion about the best kernel, therefore this study will compare 4 kernel functions, namely linear, RBF, polynomial and sigmoid.

a) *Kernel linear*: The Linear kernel was chosen because it is the simplest kernel and is used when the data is linearly overstretched.

$$K(x, y) = x \cdot y \quad (9)$$

b) *Kernel polynomial*: The Polynomial kernel was chosen because it can be used when the data is not linearly separated and is suitable for use in solving classification problems in all training data that has been normalized.

$$K(x, y) = (\gamma(x \cdot y) + C)^d \quad (10)$$

c) *Kernel Radial Basic Function (RBF)*: The RBF kernel is used when the data is not linearly separated, it is chosen because it performs well with specific parameters, and the result of the training has a small error value.

$$K(x, y) = \exp(-\gamma|x - y|^2) \quad (11)$$

d) *Kernel sigmoid*: This sigmoid kernel was chosen because it is similar to the two-layer perceptron model of the neural network, which works as an activation function for neurons.

$$K(x, y) = \tanh(\gamma(x \cdot y) + C) \quad (12)$$

3) *Evaluation*: In evaluating the performance of each SVM kernel, we implement several performance metrics, including Accuracy, Performance, Recall and F1-score. Before computing the performance metric score, we build a confusion matrix defined as an evaluation method that provides information comparing the classification of prediction results with the actual classification [27]. In a confusion matrix, there are four terms of value, namely True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). Based on these values, accuracy, precision, recall, and F1-Score values can be generated.

Accuracy is the ratio of predicted correct values of all data [28].

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \quad (13)$$

Precision indicates a correctly classified prediction of positive values divided across positive classified data [28].

$$Precision = \frac{TP}{TP + FP} \quad (14)$$

Recall compares the positive correct predicted value with the entire positive correct value [29].

$$Recall = \frac{TP}{TP + FN} \quad (15)$$

The F1-Score shows the average comparison of precision and recall values [29].

$$F1 - Score = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (16)$$

#### IV. RESULTS AND DISCUSSIONS

This stage is a decipherment of the research obtained and its explanation.

##### A. Data Preprocessing

The dataset used is the Pima Indian diabetes dataset which consists of 768 data and nine attributes. The initial stage carried out in this study is the process of collecting and processing datasets. In this study, data preprocessing was divided into three steps. The first step is the data splitting process, where the Diabetes Mellitus dataset will be divided into training data and testing data. The second step is the data normalization process to create data in the same range. The third step is an oversampling process to balance the dataset class using the SMOTE method. Data processing in the study uses the Python programming language in the google colab application.

1) *Data splitting results*: After getting the dataset, the next step is to divide the dataset into training data and testing data. The Diabetes Mellitus dataset totaled 768 data consisting of eight variables and one target/class. Then the dataset is divided into 80% training data, totaling 614 data and 20% testing data, totaling 154 data. The diagnosis of Diabetes Mellitus is divided into two, namely non-diabetics who are denoted by 0 and diabetics who are denoted by 1. Obtained diabetics totaled 268 data and non-diabetics amounted to 500 data.

2) *Data normalization results*: The normalization methods used are min-max and z-score.

Fig. 1 shows a comparison of the variables in the dataset, there are two variables being compared namely pregnancies and insulin, the data has a fairly high range of values. For example, on the insulin variable, where the value range is from 0 to over 200, this is considered unbalanced. The min-max normalization method is used to process values into the range 0-1. Fig. 2 shows the results after normalizing the min-max,

where the range of values for the insulin variable becomes smaller, ranging from 0 to 1.

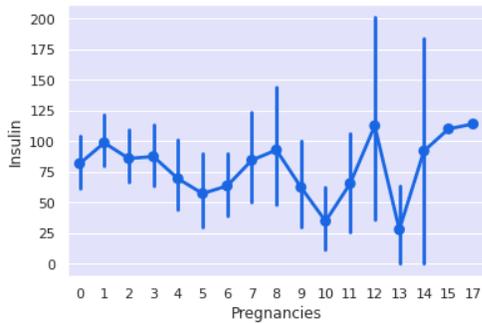


Fig. 1. Before normalization.



Fig. 2. After min-max normalization.

In addition to using the min-max method, data normalization is also carried out using the z-score method. Z-score is performed by processing the mean and standard deviation from the values of its attributes. Fig. 3 shows the results after normalizing the z-score.

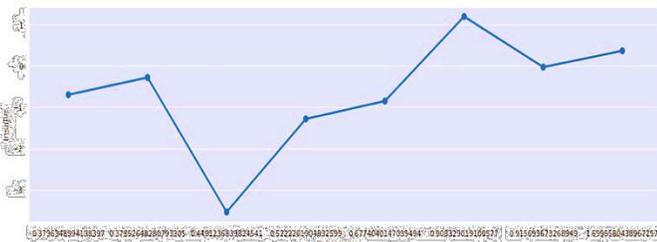


Fig. 3. After Z-score normalization.

3) *Oversampling results:* In the dataset there is a difference between the number of positive and negative classes, therefore there is a need for class balancing. Class balancing is done by oversampling using the SMOTE method and is carried out on training data only. Oversampling is carried out after splitting data so that data replication does not appear in data training and data testing [30]. It can be seen in Fig. 4, before oversampling the number of positive classes was 221 and the number of negative classes was 393. Meanwhile, after oversampling, the number between the positive and negative classes becomes the same, which is 393 so that it becomes balanced.

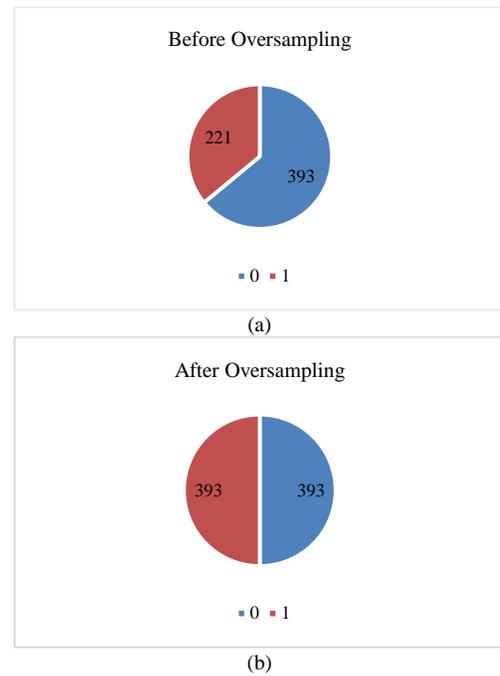


Fig. 4. (a) Data before oversampling, (b) Data after oversampling.

### B. Data Preprocessing and Evaluation

This study compared the performance of the SVM algorithm kernels for the classification of Diabetes Mellitus diseases. SVM kernels include linear, polynomial, RBF, and sigmoid kernels. Evaluation is carried out using the confusion matrix method to calculate the accuracy, precision, recall, and F1-score by optimizing the best parameters for each kernel. Each kernel on SVM has a specific parameter, the cost parameter (C) being the most commonly used value for all kernels. The gamma ( $\gamma$ ) parameter is used to determine the degree of proximity between two points to make it easier to find  $\gamma$  hyperplanes consistent with the data. The gamma parameter is used by polynomial, RBF, and sigmoid kernels. Next is the degree (d) parameter used to map data from the input space to the higher dimension space in the feature space, only the polynomial kernel uses this parameter [31]. The best parameters on the kernel are determined by trial and error. Table II evaluates the classification models of various SVM kernels before different data preprocessing is carried out.

TABLE II. RESULTS OF EVALUATION OF VARIOUS SVM KERNELS BEFORE DATA PREPROCESSING

	Kernel			
	Linear	Polynomial	RBF	Sigmoid
Accuracy	76%	77%	77%	51%
Precision	66%	68%	69%	12%
Recall	57%	55%	53%	8%
F1-Score	61%	61%	60%	10%

For this experiment in Table II, all parameter values in each kernel use auto parameters from python. The highest accuracy is obtained from the polynomial and RBF kernels, which is 77%. The highest precision was obtained from the RBF kernel, which was 69%, the highest recall was obtained from the linear kernel, which was 57% and the highest F1-score was obtained from linear and polynomial kernels, which was 61%. Table III evaluates the classification models of various SVM kernels after preprocessing data with min-max normalization and SMOTE oversampling. Meanwhile, Table IV evaluates the classification models of different SVM kernels after preprocessing data with normalization of z-score and oversampling SMOTE.

Table III and IV show that the highest accuracy is obtained by applying z-score normalization and SMOTE oversampling, which is obtained by 80% using a polynomial kernel. The polynomial kernel using the parameter value  $C=1 \gamma=0.1 d=1.5$  is obtained through trial and error to produce margin optimization values that maximize the hyperplane by mapping the data into higher dimensions. The highest precision is also obtained from the polynomial kernel, which is 65%. This shows that the higher the accuracy value, the higher the precision value. The highest recall was obtained at 79% from the RBF kernel shown in Table III. The RBF kernel uses the parameter value  $C=2.5 \gamma=1.5$ . The highest F1-score is obtained from the RBF kernel shown in Table III, which is 70%. The values in the parameters  $C, \gamma$ , and  $d$  are the most optimal values to get the maximum accuracy value. If the value is increased or decreased, the accuracy value will decrease.

TABLE III. RESULTS OF EVALUATION WITH MIN-MAX AND SMOTE

	Kernel			
	Linear	Polynomial	RBF	Sigmoid
Accuracy	77%	79%	79%	76%
Precision	61%	64%	63%	59%
Recall	72%	72%	79%	68%
F1-Score	66%	68%	70%	63%

TABLE IV. RESULTS OF VALUATION WITH Z-SCORE AND SMOTE

	Kernel			
	Linear	Polynomial	RBF	Sigmoid
Accuracy	79%	80%	77%	78%
Precision	62%	65%	61%	62%
Recall	74%	74%	72%	72%
F1-Score	68%	69%	66%	67%

## V. CONCLUSION AND FUTURE WORKS

This research produces the highest accuracy of up to 80%, obtained from polynomial kernels. So the shortcomings of previous research have been resolved in this study. By optimizing the use of the kernel on the SVM algorithm, it is proven to be able to maximize performance. Hence, it can be concluded that the SVM algorithm performs better in classifying Diabetes Mellitus.

This study found that the performance of the SVM algorithm kernel to produce the highest accuracy was obtained from the polynomial kernel. The accuracy results produced in this study can be used as an accurate and beneficial recommendation for overcoming health problems related to Diabetes Mellitus.

For further research, we can also implement other datasets that contain more data. Also, we can create different novel kernels which may gain better accuracy results. In addition, the results of this study can also be used in making applications to detect Diabetes Mellitus which can be web-based or mobile.

## REFERENCES

- N. Shamsiyah, "Mengenal diabetes melitus." In *Berdamai dengan diabetes*, pp. 1-12. Jakarta : Bumi Media, 2022.
- S. Wiyono, "Perbandingan kinerja rule zeroR dan function SMO dengan T-test dalam pengklasifikasian diagnosis penyakit diabetes mellitus." *Jurnal Teknik Elektro*, vol. 16, no. 01, pp. 23–25, 2016.
- Y. Mardi, "Data mining : Klasifikasi menggunakan algoritma C4.5." *Jurnal Edik Informatika*, vol. 2, no. 2, pp. 213–219, 2017.
- D. A. Nasution, H. H. Khotimah, and N. Chamidah, "Perbandingan normalisasi data untuk klasifikasi wine menggunakan algoritma K-NN." *CESS (Journal of Computer Engineering System and Science)*, vol. 4, no. 1, pp. 78–82, 2019.
- A. M. Argina, "Penerapan metode klasifikasi k-nearest neighbor pada dataset penderita penyakit diabetes." *Indonesian Journal of Data Sciencs*, vol. 1, no. 2, pp. 29–33, 2020.
- Noviandi, "Implementasi algoritma decision tree C4.5 untuk prediksi penyakit diabetes." *Jurnal INOHIM*, vol. 6, no. 1, pp. 1–5, 2018.
- M. Azhari, Z. Situmorang, and R. Rosnelly, "Perbandingan akurasi, recall, dan presisi klasifikasi pada algoritma C4.5, random forest, SVM dan naive bayes." *Jurnal Media Informatika Budidarma*, vol. 5, no. 2, pp. 640–651, 2021.
- Y. Zhao, C. Zhang, Y. Zhang, Z. Wang, and J. Li, "A review of data mining technologies in building energy systems: Load prediction, pattern identification , fault detection and diagnosis." *Energy and Built Environment*, vol. 1, no. 2, pp. 149–164, 2020.
- U. P. Harapan, D. E. Ratnawati, and A. W. Widodo, "Klasifikasi penyakit gigi dan mulut menggunakan metode support vector machine." *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, vol. 2, no. 2 2018.
- A. F. Rina Kurniasari, "Penerapan algoritma C4.5 untuk penjurusan siswa sekolah menengah atas." *Jurnal Ilmiah Komputer dan Informatika (KOMPUTA)*, vol. 8, no. 1, 2019.
- B. A. H. and F. A. S. B. Helmi Imaduddin, "Arison of support vector machine and decision tree methods in the classification of breast cancer." *Jurnal Pendidikan Teknologi Informasi*, vol. 5, pp. 22–30, 2021.
- R. A. Helena Nurramdhani Irmanda, "Klasifikasi jenis pantun dengan metode support vector machines (SVM)." *JURNAL RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 4, no. 5, pp. 915–922, 2020.
- D. M. Ahmad Harmain, Paiman, Henri Kurniawan, Kusri, "Normalisasi data untuk efisiensi k-means pada pengelompokan wilayah berpotensi kebakaran hutan dan lahan berdasarkan sebaran titik panas." *TEKNIMEDIA*, vol. 2, no. 2, pp. 83–89, 2021.
- H. E. Wahanani, M. H. P. Swari, and F. A. Akbar, "Case based reasoning prediksi waktu studi mahasiswa menggunakan metode euclidean distance dan normalisasi min-max." *Jurnal Teknologi Informasi dan Ilmu Komputer*, vol. 7, no. 6, pp. 1279–1288, 2020.
- R. Fatwa, I. Cholissodin, and Y. A. Sari, "Penerapan metode extreme learning machine untuk prediksi konsumsi batubara sektor pembangkit listrik tenaga uap." *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, vol. 3, no. 11, pp. 10749–10755, 2019.
- U. Al, A. Mandar, and S. Basri, "Novelty ranking approach with z-score and fuzzy multi-attribute decision making combination." *International Journal of Engineering & Technology*, vol. 7, no. 7, pp. 476–480, 2018.

- [17] T. M. Fahrudin, P. A. Riyantoko, K. M. Hindrayani, and M. H. P. Swari, "Cluster analysis of hospital inpatient service efficiency based on BOR, BTO, TOI, AvLOS indicators using agglomerative hierarchical clustering." *Telematika*, vol. 18, no. 2, p. 194, 2021.
- [18] R. Siringoringo, "Klasifikasi data tidak seimbang menggunakan algoritma SMOTE dan k-nearest neighbor." *Journal Information System Development (ISD)*, vol. 3 no. 1, 2018.
- [19] N. Nurajjiah, D. A. Ningtyas, and M. Wahyudi, "Klasifikasi siswa SMK berpotensi putus sekolah menggunakan algoritma decision tree, support vector machine dan naïve bayes." *Jurnal Khatulistiwa Informatika*, vol. 7, no. 2, 2019.
- [20] H. Nalatissifa, W. Gata, S. Diantika, and K. Nisa, "Perbandingan kinerja algoritma klasifikasi naive bayes, support vector machine (SVM), dan random forest untuk prediksi ketidakhadiran di tempat kerja." *Jurnal Informatika Universitas Pamulang*, vol. 5, no. 4, p. 578, 2021.
- [21] A. A. Kasim and M. Sudarsono, "Algoritma support vector machine (SVM) untuk klasifikasi ekonomi penduduk penerima bantuan pemerintah di kecamatan simpang raya sulawesi tengah." *SEMNASSTIK*, pp. 568–573, 2019.
- [22] S. A. Naufal, A. Adiwijaya, and W. Astuti, "Analisis perbandingan klasifikasi support vector machine (SVM) dan k-nearest neighbors (KNN) untuk deteksi kanker dengan data microarray." *JURIKOM (Jurnal Riset Komputer)*, vol. 7, no. 1, pp. 162–168, 2020.
- [23] S. Widodo, R. N. Rohmah, B. Handaga, L. Dyah, and D. Arini, "Lung diseases detection caused by smoking using support vector machine." *TELEKOMUNIKA*, vol. 17, no. 3, pp. 1256–1266, 2019.
- [24] R. Wati and S. Ernawati, " Analisis sentimen persepsi publik mengenai PPKM pada twitter berbasis SVM menggunakan python." *Jurnal Pendidikan Teknologi Informasi*, vol. 06, pp. 240–247, 2021.
- [25] E. Anindika Sari, M. Thereza Br. Saragih, I. Ali Shariati, S. Sofyan, R. Al Baihaqi, and R. Nooraeni, "Klasifikasi kabupaten tertinggal di kawasan timur indonesia dengan support vector machine." *JIKO (Jurnal Informatika dan Komputer)*, vol. 3, no. 3, pp. 188–195, 2020.
- [26] R. H. Muhammadiyah, T. G. Laksana, and A. B. Arifa, "Combination of support vector machine and lexicon-based algorithm in twitter sentiment analysis." *Khazanah Informatika: Journal Ilmu Komputer dan Informatika*, vol. 8 no. 1, 2021.
- [27] M. Syukron, R. Santoso, and T. Widiariyah, "Perbandingan metode smote random forest dan smote xgboost untuk klasifikasi tingkat penyakit hepatitis C pada imbalance class data." *Jurnal Gaussian*, vol. 9, no. 3, pp. 227–236, 2020.
- [28] M. Rangga, A. Nasution, and M. Hayaty, " Perbandingan akurasi dan waktu proses algoritma K-NN dan SVM dalam analisis sentimen twitter." *Jurnal Informatika*, vol. 6, no. 2, pp. 212–218, 2019.
- [29] A. Ridhovan, A. Suharso, "Penerapan metode residual network (RESNET) dalam klasifikasi penyakit pada daun gandum." *JUPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)*, vol. 7, no. 1, pp. 58–65, 2022.
- [30] A. A. Arifiyanti and E. D. Wahyuni, "Smote : Metode penyeimbangan kelas pada klasifikasi data mining." *SCAN-Jurnal Teknologi Informasi dan Komunikasi*, vol. 15 no. 1, pp. 34–39, 2020.
- [31] I. M. Yulietha and S. Al Faraby, "Klasifikasi sentimen review film menggunakan algoritma support vector machine." *eProceedings of Engineering*, vol. 4, no. 3, pp. 4740–4750, 2017.

# Deep Study of CRF Models for Speech understanding in Limited Task

Marwa Graja

College of Computer and Information Sciences  
Jouf University, KSA

**Abstract**—In this paper, we propose to evaluate in depth CRF models (Conditional Random Fields) for speech-understanding in limited task. To evaluate these models, we design several models that differ according to the level of integration of local dependencies in the same turn. As we propose to evaluate these models on different types of processed data. We perform our study on a corpus where turns are not segmented into utterances. In fact, we propose to use the whole turn as one unit during training and testing of CRF models. This represents the natural way of conversation. The language used in this work is the Tunisian Arabic dialect. The obtained results prove the robustness of CRF models when dealing with raw data. They are able to detect the semantic dependency between words in the same speech turn. Results are important when CRF models are designed to take into account the words with deep dependencies in the same turn and with advanced preprocessed data.

**Keywords**—Speech understanding; Arabic dialect; CRF models

## I. INTRODUCTION AND RELATED WORKS

Spoken Language Understanding is an important component in spoken dialogue systems. It aims to extract concepts from an utterance to clarify speech meaning. Therefore, the key link in an automatic understanding process revolves around the correspondence between the set of words in the utterance and the set of semantic concepts. In order to resolve this correspondence, the first research works in this field exploited linguistic formalisms such as regular grammars and context-free grammars. Recent works have rather oriented towards the exploitation of machine learning models for concept detection, these models are widely used for the semantic annotation of speech utterances.

Our overview of the literature showed that learning models constitute the dominant context for speech understanding due to the performances recorded particularly in restricted domains. These models enjoy several advantages reported by [1] [2] [3]. Indeed, the intervention of a human expert is limited to the labeling of data, which represents an easier task than the modeling of grammars or patterns. Moreover, these models offer better portability since they are domain and language independent. However, the effectiveness of these models is sensitive to the used corpus, which must be representative and large, in order to determine their parameters [4].

Machine learning models are classified into generative and discriminative models [2] and they are widely applied to speech understanding. HMM (Hidden Markov Models) is an example of generative models and they are used by [5] for speech understanding of Spanish language, using the DIHANA

corpus. The DIHANA corpus task deals with requests of information about railway services. This work uses HMM in the most realistic situations where dialogues are not segmented into utterances. The results of their work are very important. They obtained 92% as F-measure. This good result is due to the large size of the corpus used for training models.

In the literature, several studies show that discriminant models perform better than generative [6] [2] [7]. CRF models (Conditional Random Fields), as an example of discriminant models, have been widely exploited in many tasks in natural language processing such as semantic annotation and syntactic analysis [8], [9] [10]. A particular distinction is reported for CRF models whose performance exceeded that of other models, [11] [2] [12]. It is so important to notice that the CRF models have the capacity to integrate correlated characteristics that make it possible to take into account the local context of an utterance. All of these observations encouraged us to exploit these models in the context of the speech understanding of the Arabic dialect.

Several works have shown the robustness of Conditional Random Fields (CRF) models for request information in the French language using the MEDIA corpus [12]. The MEDIA corpus is manually annotated with semantic concepts of touristic information. Turns in this corpus are segmented into utterances, which simplifies the speech understanding. Raymond et al. [12] have used CRF models and domain knowledge through a set of rules made manually. This has reduced the conceptual error rate (from 11.2% to 10.9% as CER), and has increased the performance of the system to 92% as F-measure. This justifies the advantage of segmenting turns into utterances and the important size of the training corpus.

In addition, CRF models offer two major advantages. On the one hand, they allow segmentation and conceptual annotation taking into account the local context of the utterance. On the other hand, they make it possible to guarantee convergence towards the most probable concepts by taking into account all the previous and following observations in the statement [13]. Indeed, these models have the ability to use all the observations of a sequence to predict a conceptual label. This represents an interesting distinction compared to HMM (Hidden Markov Models).

In this paper, we propose to evaluate the performance of CRF models. We designed several models that differ according to the level of integration of local dependencies in the same turn. We also propose to use several processing levels on the corpus. In addition, almost all learning-based understanding

methods have been interested in modeling speech turns segmented into utterances, we suggest to use turn as a whole unit (not segmented into utterances) to test the performance of CRF models, which represents the natural way of conversation.

This paper is organized as follows. In Section II, we present CRF models for speech understanding. Section III presents Spoken Dialogue Corpus for Tunisian Dialect. Section IV deals with evaluation metrics. In Section V, we present experiments and discussion. A conclusion is drawn in Section VI.

## II. CRF MODELS FOR SPEECH ANNOTATION

Conditional Random Fields (CRFs), initiated by Lafferty, are discriminant models that define the conditional probability of observation sequences according to label sequences [14]. Lafferty defines the conditional sequence labeling probability  $Y=y_1...y_n$  given an observation sequence  $X=x_1...x_n$  as follows:

$$P(Y|X) = \frac{1}{z(X)} \exp(\sum_j \lambda_j t_j(y_{i-1}, y_i, X, i) + \sum_k \mu_k s_k(y_i, X, i)) \quad (1)$$

With

$$z(X) = \sum_y \exp(\sum_j \lambda_j t_j(y_{i-1}, y_i, X, i) + \sum_k \mu_k s_k(y_i, X, i)) \quad (2)$$

- $z(X)$  is a factor that normalizes the probabilities.
- $t_j(y_{i-1}, y_i, X, i)$  represents the transition characteristic function for an observation sequence between the labels at position  $i$  and  $i-1$ .
- $s_k(y_i, X, i)$  represents the characteristic function of the state of the label for a sequence of observations .
- $\lambda_j$  et  $\mu_k$  are real values which make it possible to attribute a weight to each characteristic function to specify its importance. These values therefore make it possible to characterize the discriminating power of the model. These parameters are fixed during the learning phase and make it possible to maximize the likelihood on a set of already annotated data.

Referring to the model defined in Eq. (1), the most likely sequence of concepts for labeling a sequence of input words is:

$$Y^* = \arg \max_y P(Y|X) \quad (3)$$

CRF is modeled by undirected graph models (see Fig. 1) to define a probability distribution over a process of label  $Y$  given an observation  $X$ , by maximizing a conditional probability [14].

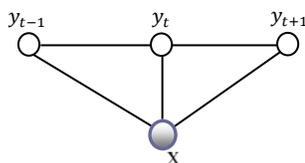


Fig. 1. Graphical design for CRF models

In this graph, the set of vertical nodes are two random fields  $X$  and  $Y$ , respectively describing the set of observations and the set of annotations. Two variables linked in the graph

express that one depends on the other. Based on this, each node  $y_t$  depends on the preceding node  $y_{t-1}$  and the following node  $y_{t+1}$ , and implicitly on the variable  $x$ . Therefore, each variable  $y_t$  must be linked to a variable  $x$  to guarantee the dependence between the labels on the one hand and the sequences of observations on the other.

Learning CRF models consists in determining from the learning corpus, the vector  $\theta = \{\lambda_1, \lambda_1, \dots, \lambda_{k1}, \mu_1, \mu, \dots, \mu_{k2}\}$  that represents the weight vector of the characteristic functions  $t$  and  $s$ . After the learning step, the exploitation of CRF models on new data consists in finding the most probable sequence of states given a new sequence of observations, which are not encountered in the training corpus. We perform this by applying the Viterbi algorithm, as it is the case with HMM models.

## III. SPOKEN DIALOGUE CORPUS FOR TUNISIAN DIALECT – TUDICOI

### A. Corpus Description

The TUDICOI corpus, used in this work, consists of spontaneous oral dialogues of railway request information, in Tunisian Dialect (TD). The purpose of these requests is to consult the timetables of the train, the type of train, the destination of the train, the route taken by the train, the price and types of tickets and the reservation of tickets. We should notify that several requests can be combined during a dialogue between tellers and customers [15].

The transcribed part of the TuDiCoI corpus consists of 1825 dialogues representing 12182 turns. These turns consists of 6533 customer turns and 5649 agent turns. We list the features of the TuDiCoI corpus in Table I.

TABLE I. SPECIFICATION OF THE TUDICOI CORPUS

# dialogues	1 825
# customer turns	6 533
# agent turns	5 649
# customer word	21 551
# customer vocabulary	1 437

Note that on average, each dialogue consists of three turns for the customer and three turns for the agent. Additionally, each customer turn is comprised of an average of 3.3 words. It is important to note that this average is low due to the agglutinative aspect of words in the dialect and the frequent use of keywords to request for information.

TABLE II. CHARACTERISTICS OF SOME DIALOGUE CORPORA IN LIMITED TASKS

Corpus	#D	#T	#V	#W	A	L	Task
Trains 93	98	5 900	860	55 000	H-H	Ang	Manufacture and shipment of goods on the railways.
DIHANA	900	15 413	823	48 243	H-M	Esp	Railway information.
SARF	350	9 763	827	117 156	H-M	Ar	Railway information.

TuDiCoI	1 825	12 182	1 437	21 551	H-H	TD	Railway information.
TARIC	4 662	18 657	--	71 684	H-H	TD	Railway information.
MEDIA	1 257	38 434	2 715	156 048	H-M	Fr	Hotels reservation.

Table II shows the characteristics of different dialogue corpora used in other projects in different languages. It should be noted that #D designates the number of dialogues, #T the number of speaking turns, #V the size of the vocabulary, #W the number of words. A designates the type of corpus (H-H for Human-Human and H-M for Human-Machine). Finally, L provides information on the language used (Eng. for English, Esp. for Spanish, Ar. for Arabic, TD for the Tunisian dialect and Fr. for French). These corpora vary in size from a few tens to thousands of dialogues.

### B. Annotation Schema

We proposed in this work an annotation scheme to perform the manual concept annotation step. Table III summarizes the annotation scheme defined for dialogue acts and semantic concepts [16].

TABLE III. SEMANTIC CONCEPTS

Domain concepts		Requests concepts	
Train	Ticket_Numbers	Path_Req	Existence_Req
Train_Type	Ticket	Hour_Req	Trip_timeReq
Departure_hour	Hour_Cpt	Price_Req	Clarification_Req
Arrival_hour	Departure_Cpt	Dialogue concepts	
Day	Arrival_Cpt	Rejection	Salutation_Begin
Origin	Price_Cpt	Acceptance	Salutation_End
Destination	Class_Cpt	Politeness	
Fare	Trip_time	Out of vocabulary	
Class	Ticket_type	Out	
Link concepts			
Choice			
Coordination			

Due to the complexity of the annotation task, effort, and manual verification, we have annotated only 1476 dialogues which represents 5047 customer turns. The characteristics of the annotated corpus are summarized in Table IV.

TABLE IV. CHARACTERISTICS OF THE ANNOTATED PART OF THE TUDICOI CORPUS

# Annotated dialogues	1 476 / 1 825
# Customer annotated turns	5 047 / 6 533
# Annotated Customer words	16 772 / 21 551

In order to define the parts of the TuDiCoI corpus used for our evaluations, we have divided the annotated corpus into two parts. The first part of the corpus is used for learning and it represents about 80% of the total size, while the second part constitutes 20% of the corpus used for the test. Table V provides information on the characteristics of these two different parts in terms of number of dialogues, speaking turns and words.

TABLE V. CHARACTERISTICS OF TRAINING AND TEST CORPUS

	Training	Test
Dialogues	1202	267
Turns	4131	906
Words	13555	3217

Since we are interested in literal understanding, which does not depend on dialogical context, we have classified all the speaking turns of the test part into three types, according to the recommendation proposed by the ARPA community [17], namely sets A, D and X. Table VI presents the characteristics of these different sets. This classification makes it possible to give an overview of the types of turns contained in the test part.

TABLE VI. CHARACTERISTICS OF THE TEST CORPUS DIVIDED INTO THREE SETS A, D AND X

# Client turns			
A	D	X	Total
379	482	45	906
41.83%	53.21%	4.96%	100%

The first set corresponds to the context-independent customer turns (Set A). This set contains the turns that have no connection with the history of the dialogue. While the second set corresponds to the context-dependent ones (Set D). This set contains the turns that have a relationship with the dialogical context. The third set corresponds to out-of-context turns of the dialogue (Set X). This set includes marginal turns that are not related to the domain. Table VII shows an example for each series.

TABLE VII. EXAMPLE FOR EACH SET A, D AND X

Set	Transcription ◀ / ▶ Transliteration / Translation
A	تكية لتونس برميير كلاس / tikiyh ltuwnis brumiyir klaAs / A first class ticket to Tunis
D	بقداش هي ؟ / b.qad~Aš hiya ? / How much does it cost ?
X	دخلت عالانترنت وشفت / dxalt ʕalAantirnaAt wuʕuft / I am connected to the Internet and I saw

We utilize these different sets of the TuDiCoI corpus for the evaluation of CRF-based speech understanding method for the TD.

Almost all speech-understanding methods are interested in modeling speech turns segmented into utterances. The alternative we have proposed is to use the turn as a whole unit for training and testing the performance of CRF models [16]. This represents the natural way of conversation.

### C. Pretreatments

To evaluate CRF-based speech understanding, we prepared three versions of the TuDiCoI corpus:

- The first version (version I) is a raw version which is not pre-processed, thus increasing the complexity of the structure of the dialect turns. In this version, the words do not respect the spelling transcription guide. Therefore, a word can be written in different spellings. Likewise, this version has morphological problems

such as the agglutination of a particle with the word, which follows it. The evaluations carried out on this version of the corpus is used to test the performance of the CRF models on data not processed in advance.

- The second version of the annotated corpus (version II), is partially preprocessed. This version has undergone spelling correction, morphological analysis of verbs and nouns, as well as synonymy analysis processing.
- The third version of the annotated corpus (version III) presents an improvement compared to the second version and which consists in processing the agglutinations of the names of cities, which makes it possible to dissociate the particle, if it exists, from the name that is attached to it.

#### D. Tabular Corpus

After the manual labeling step into concepts, we converted each version of the annotated corpus into a standard representation adopted by CRF models. This representation uses a set of labels called the BIO notation (Begin Inside Outside) [18], in which:

- The label starting with "B-???" indicates the beginning of the conceptual segment.
- The label "I-???" denotes any meaningful word that is part of the conceptual segment.
- The label "O" is assigned for words that do not refer to any conceptual label.

سلام	B- Salutation_Begin
عليكم	I- Salutation_Begin
تذكرة	B-Ticket
لتونس	B-Destination
بقداه	B-Price_Req
زوز	B-Ticket_Numbers
زوز	B-Ticket_Numbers
لتونس	B-Destination
وقتاش	B-Dep_Hour_Req
يخرج	B-Departure_Cpt
هو	O
تسكرة	B-Ticket
لتونس	B-Destination
ماضي	B-Dep_hour
ساعة	I-Dep_hour

Fig. 2. Example of BIO notation from the TUDICOI corpus

The advantage of using BIO notation is that it is able to segment a set of words into several conceptual segments and display them one after the other [19] [7]. An example from TUDICOI corpus in BIO notation is shown in Fig. 2.

The tabular corpus is used for training and testing CRF models. Conceptual labeling using CRF models consists in finding the best sequence of states, given a sequence of input observations. This problem is solved using the Viterbi algorithm due to the linear topology of CRF models [20]. This algorithm makes it possible to give the list of n best results.

#### IV. EVALUATION METRICS

The evaluation makes it possible to evaluate the conceptual correspondence, which consists in seeking the pairing between a set of words of a turn and a set of semantic concepts. For this, we use the Concept Error Rate (CER). The CER makes it possible to compare the list of reference semantic concepts with the list of concepts emitted by the system according to the following equation:

$$CER = \frac{\# \text{ incorrect predicted concepts}}{\# \text{ reference concepts}} \quad (4)$$

We use other measures to evaluate conceptual labelling such as Precision, Recall, and F-measure. The Precision represents the number of correct concepts found compared to the number of concepts found by the system.

$$Precision = \frac{\# \text{ correct concepts found}}{\# \text{ concepts found}} \quad (5)$$

The Recall represents the number of correct concepts found by the system with regard to the reference concepts.

$$Recal = \frac{\# \text{ correct concepts found}}{\# \text{ reference concepts}} \quad (6)$$

The F-measure combines Precision and Recall according to the following equation:

$$F - \text{measure} = \frac{2 \times \text{Recal} \times \text{Precision}}{\text{Recal} + \text{Precision}} \quad (7)$$

We used in our experiments, the free tool CRF++ for the training and testing steps. It should be noted that the CRF++ tool implements learning by Newtonian method and uses decoding using the Viterbi algorithm.

#### V. RESULTS AND DISCUSSION

In order to evaluate the performance of CRF models, we used several models that differ according to the level of integration of local dependencies in the same turn. These dependencies vary according to the unigram (one word) or bigram (two words) interval of the word to label.

After an initial test phase, we limited the number of models tested to four.

- The first (Model 0) is a model that does not take into account any dependence between the words of the same turn. In this case, CRF models play the role of a simple semantic tagger.
- The second (Model 1) is a model that uses a two-word window taking into account the previous word and the next word in the same turn.
- The third (Model 2) is a model that uses a window involving two words before and two words after the current word.
- The fourth model (Model 3) consists of improving the third model by adding two local dependencies. This dependency uses two bigrams taking into account the current word with the precedent word (respectively with the next word).

Then, we use these models for learning the CRF parameters based on different versions of the annotated corpus (version I,

II and III). The Table VIII, Table IX and Table X illustrate the results of the evaluation of the concept labeling in terms of Precision, Recall, F-measure and Concept Error Rate (CER).

TABLE VIII. RESULTS OF THE CONCEPTUAL LABELING OF VERSION I OF THE CORPUS

		Model 0	Model 1	Model 2	Model 3
<b>CER</b>	(%)	20.05	10.44	8.98	8.47
<b>Precision</b>	(%)	79.19	88.50	89.86	90.33
<b>Recall</b>	(%)	76.36	80.45	79.68	79.21
<b>F-measure</b>	(%)	77.75	84.28	84.46	84.40

TABLE IX. RESULTS OF THE CONCEPTUAL LABELING OF VERSION II OF THE CORPUS

		Model 0	Model 1	Model 2	Model 3
<b>CER</b>	(%)	18.65	9.05	8.39	7.70
<b>Precision</b>	(%)	80.72	90.22	90.75	91.45
<b>Recall</b>	(%)	78.13	83.57	82.47	82.47
<b>F-measure</b>	(%)	79.40	86.77	86.41	86.73

TABLE X. RESULTS OF THE CONCEPTUAL LABELING OF VERSION III OF THE CORPUS

		Model 0	Model 1	Model 2	Model 3
<b>CER</b>	(%)	20.05	9.19	8.51	7.86
<b>Precision</b>	(%)	79.28	90.12	90.66	91.26
<b>Recall</b>	(%)	76.75	83.93	82.68	82.24
<b>F-measure</b>	(%)	77.99	86.92	86.48	86.52

Based on these experiments, we notice that the CER decreases with the improvement of the quality of data used for learning. This clearly shows that the pre-processing carried out makes it possible to improve the speech understanding. We noticed also that the models that take into account the dependence between the different words of the same turn (Model 2 and Model 3) make it possible to improve the speech understanding. This is justified by the decrease in the CER and the increase in the F-measure.

Besides these results, we justify the robustness of CRF models with not processed data. Table VIII shows that the F-measure is 77.75% for the "Model 0" which does not take into account underlying dependencies, and reaches 84.40% for the "Model 3" by introducing the bi-model gram.

The examination of the errors made by the CRF models, directed us to carry out other experiments by exploiting the same test corpus, but with considering the dependence of the turn according to the dialogical level and exploiting the sets A, D and X (Table VI). Based on this classification, we tested the CRF models using these different sets on the different versions of preprocessing corpus.

Indeed, we obtained three different versions according to the processing performed for each set A, D and X, starting from the raw version to the fully processed version for each series. The conceptual labeling of the different series is based

on the same CRF models based on the different models (Model 0, Model 1, Model 2 and Model 3) presented previously.

Table XI summarizes the results obtained in terms of CER, Precision, Recall and F-measure. From these results, we notice that the CER obtained on the type A speech turns (speech context-independent set) is the lowest rate, comparing it with the results of sets D and X. Therefore, we can conclude that a large part of the errors is due to the presence of out-of-context statements of the dialogue (Set X) and of context-dependent statements (Set D). These results are expected since we are interested in this work in the literal understanding, which does not depend on the dialogical context, so, the turns depending on the context increase the error.

There are other sources contribute to the increase CER. It is mainly about the appearance of terms that are not processed in the training corpus. This is due to the presence of certain phenomena linked to spontaneous speech such as hesitation allowing the addition of out-of-vocabulary words.

TABLE XI. RESULTS OF THE CONCEPTUAL LABELING OF SETS A, D AND X WITH DIFFERENT TYPES OF PROCESSED DATA (VERSION I, II AND III)

		Version I		
		A	D	X
<b>Model 0</b>	CER (%)	16.93	23.21	38.59
	Precision (%)	82.58	75.68	59.25
	Recall (%)	80.32	72.24	56.14
	F-measure (%)	81.43	73.92	57.65

<b>Model 1</b>	CER (%)	8.83	12.53	12.28
	Precision (%)	90.45	85.99	84.09
	Recall (%)	83.73	76.95	64.91
	F-measure (%)	86.96	81.22	73.26

<b>Model 2</b>	CER (%)	8.03	10.26	10.52
	Precision (%)	91.17	88.15	85.00
	Recall (%)	82.99	76.36	59.64
	F-measure (%)	86.89	81.83	70.10

<b>Model 3</b>	CER (%)	7.29	9.92	10.51
	Precision (%)	91.87	88.44	86.36
	Recall (%)	82.46	75.94	66.66
	F-measure (%)	86.91	81.71	75.24

		Version II		
		A	D	X
<b>Model 0</b>	CER (%)	15.38	21.76	42.10
	Precision (%)	84.20	77.36	55.55
	Recall (%)	82.00	74.37	52.63

	F-measure (%)	83.09	75.83	54.05
--	---------------	-------	-------	-------

<b>Model 1</b>	CER (%)	7.02	11.34	17.54
	Precision (%)	92.55	87.53	78.26
	Recall (%)	87.35	79.66	63.15
	F-measure (%)	89.88	83.41	69.90

<b>Model 2</b>	CER (%)	6.95	10.25	14.03
	Precision (%)	92.49	88.55	81.39
	Recall (%)	85.75	79.23	61.40
	F-measure (%)	88.99	83.68	70.00

<b>Model 3</b>	CER (%)	6.15	9.66	8.77
	Precision (%)	93.30	89.13	87.17
	Recall (%)	85.81	79.24	59.64
	F-measure (%)	89.40	83.89	70.83

		<b>Version III</b>		
		<b>A</b>	<b>D</b>	<b>X</b>
<b>Model 0</b>	CER (%)	17.61	22.23	42.10
	Precision (%)	81.92	76.76	55.55
	Recall (%)	79.84	73.74	52.63
	F-measure (%)	80.87	75.22	54.05

<b>Model 1</b>	CER (%)	6.84	11.95	17.54
	Precision (%)	92.75	86.97	78.26
	Recall (%)	87.73	73.74	63.15
	F-measure (%)	90.17	75.22	69.90

<b>Model 2</b>	CER (%)	6.65	10.95	10.52
	Precision (%)	92.83	87.82	85.00
	Recall (%)	86.23	79.01	59.64
	F-measure (%)	89.41	83.18	70.10

<b>Model 3</b>	CER (%)	5.74	10.61	8.77
	Precision (%)	93.76	88.10	87.17
	Recall (%)	86.30	78.67	59.64
	F-measure (%)	89.87	83.12	70.83

In conclusion, CRF models perform well even with not processed turns. On the other hand, conceptual labeling based on CRF models failed when dealing with new terms that are not in the training corpus. These terms can be non-vocabulary

words or domain words. This last case is mainly due to the reduced size of the corpus used to learn the CRF models.

## VI. CONCLUSION

In this work, we proposed to evaluate in depth the performance of CRF models in the context of speech understanding in dialogue systems. We tested CRF in different models and in different types of processed data. We proved that these models show robustness against noisy data. They recorded good results for conceptual labeling (F-measure of 86.52%). Thus, we found that CRF models have the ability to detect task-specific compound words and label them correctly. These interpretations confirm the performance of these models even for under-resourced languages. As future work, we planify to compare these results with deep learning models for the same task to compare performance between machine learning using CRF models and deep learning models such as CNN (Convolutional Neural Network).

## REFERENCES

- [1] W. Minker, "Compréhension Automatique de la Parole Spontanée". Éditions de l'Harmattan, 1999.
- [2] C. Raymond, and G. Riccardi, "Generative and Discriminative Algorithms for Spoken Language Understanding. In proceeding of Interspeech", pp.1605-1608, 2007.
- [3] S. Young, M. Gasic, B. Thomson, and J. D. Williams, "POMDP-Based Statistical Spoken Dialog Systems: A Review". In proceeding of the IEEE, vol. 101, pp.1160 1179, 2013.
- [4] C. Lhioui, A. Zouaghi, and M. Zrigui, "A Combined Method Based on Stochastic and Linguistic Paradigm for the Understanding of Arabic Spontaneous Utterances". In proceeding of the international conference on intelligent text processing and computational linguistics (CICLing), lecture notes in computer science (LNCS) springer, vol. 7817, pp.549-558, 2013.
- [5] C. D. Martínez-Hinarejos, B. José-Miguel, and G. Ramón, "Statistical framework for a Spanish spoken dialogue corpus," Speech Commun., vol. 50, pp. 992–1008, 2008.
- [6] Y. Y. Wang and A. Acero, "Discriminative Models for Spoken Language Understanding". In proceeding of international conference on spoken language processing (ISCA), pp.1766 1769, 2006.
- [7] D. Y. Liliana and C. Basaruddin, "A review on conditional random fields as a sequential classifier in machine learning", 2017 International Conference on Electrical Engineering and Computer Science (ICECOS), Palembang, Indonesia, 2017, pp. 143-148.
- [8] Y. Tsuruoka, J. Tsujii, and S. Ananiadou, "Fast Full Parsing By Linear-Chain Conditional Random Fields". In proceeding of the conference of the European chapter of the ACL, pp.790–798, 2009.
- [9] I. Tellier, I. Eshkol, S. Taala, and J.P. Prost, "Pos-tagging for oral texts with CRF and category decomposition". Research in computing science, vol. 46, pp.79–90, 2010.
- [10] A. Barhoumi, C. Aloulou, I. Zitouni, and L.H. Belguith, "Analyse Syntaxique Statistique de la Langue Arabe ". (CEC-TAL 2015) pp.38-46, 2015.
- [11] Y.Y. Wang, A. Acero, M. Mahajan, J. Lee, "Combining Statistical and Knowledge-Based Spoken Language Understanding in Conditional Models". In proceeding of the international conference on computational linguistics and the annual meeting of the association for computational linguistics COLING-ACL, pp.882-889, 2006.
- [12] S. Hahn, M. Dinarelli, C. Raymond, F. Lefèvre, O. Lehnén, R. De Mori, A. Moschitti, H. Ney, and G. Riccardi, "Comparing Stochastic Approaches to Spoken Language Understanding in Multiple Languages". IEEE transaction on audio, speech, and language processing, vol. 19, pp.1569-1583, 2011.
- [13] A. Deoras, G. Tur, R. Sarikaya, and D. Hakkani-Tür, "Joint Discriminative Decoding of Words and Semantic Tags for Spoken

- Language Understanding”. IEEE transactions on audio, speech, and language processing, vol. 21, pp.1612-1621, 2013.
- [14] J. Lafferty, A. McCallum, F. Pereira, “Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data”. In proceeding of the international conference on machine learning (ICML), pp.282-289, 2001.
- [15] M. Graja, M. Jaoua, and L.H. Belguith, “Statistical Framework with Knowledge Base Integration for Robust Speech Understanding of the Tunisian Dialect”. IEEE Transactions on Audio, Speech, and Language Processing (T-ASLP), 2015.
- [16] M. Graja, M. Jaoua, and L.H. Belguith, “Discriminative Framework for Spoken Tunisian Dialect Understanding”. In proceeding of international conference on statistical language and speech processing (SLSP). Lecture notes in computer science, LNCS, vol. 7978, pp.102-110, Springer, 2013.
- [17] W. Minker and S. Bennacef S, “Speech and Human Machine Dialog”. The springer international series in engineering and computer science, kluwer academic publishers, vol. 770, 2004.
- [18] C. Sutton and A. McCallum, “An Introduction to Conditional Random Fields. Foundations and trends in machine learning”, vol. 4, pp.267–373, 2012.
- [19] C. Sutton and A. McCallum, “An Introduction to Conditional Random Fields for Relational Learning”, MIT Press, 2006.
- [20] T. Do and T. Artières, “Champs de Markov Conditionnels pour le Traitement de Séquences “. Revue des nouvelles technologies de l’information, extraction et gestion de connaissances, vol. 2, pp.639-650, 2006.

# Paw Search-A Searching Approach for Unsorted Data Combining with Binary Search and Merge Sort Algorithm

## Paw Search Algorithm

Md. Harun Or Rashid<sup>1</sup>, Ahmed Imtiaz<sup>2</sup>

Department of Computer Science and Engineering, Rangamati Science and Technology University, Jhagrabil, Rangamati, 4500,  
Chattogram, Bangladesh<sup>1,2</sup>

**Abstract**—Searching is one of the oldest core mechanism of nature. Nature is changing gradually along with searching approaches too. Data Mining is one of the most important industrials topic now-a-days. Under this area all social networks, governmental or non-governmental institutions and ecommerce industries produce a huge number of unsorted data and they are to utilize it. For utilizing this huge number of unsorted data there needs some specific features based unsorted data structure tools like searching algorithm. At present there are several sorted data based searching algorithms like Binary Search, Linear Search, Jump Search and Interpolation Search and so on. In this paper of Paw Search Algorithm, it is fully focused to develop a new approach of searching that can work on unsorted data merging several searching techniques and sorting techniques. This algorithm starts its operation by breaking down the given unsorted array into several blocks by making the square root of the length of the given array. Then these blocks will be searched within its specific formula till the target data is found or not, and in the inner side of each block there will be performed Merge Sort and Binary Search approach gradually. Time and Space Complexity of this Paw Search algorithm is comparatively optimal.

**Keywords**—Paw; search; unsorted; data; blocks; square root

### I. INTRODUCTION

In this present world, technology is the heart of all activities, operations and so on, and the large amount of unsorted data sets generated from different sites of the world as well as different institutions are the largest and best resources of the present technology. Managing this large amount of data with proper data structure techniques is the best tool for leading this IT world now-a-days. In this paper, we will go through a new technique of searching of data from the given unsorted array of data. There are several techniques raised now-a-days, but here we will go through a new dimension of searching and merging several built in techniques along with some new approaches to generate an optimal output.

In present world there are tons of unsorted data produced within a minimal time randomly. There are several searching algorithms like Linear Search [1, 2, 3], Binary Search [1, 2], Jump Search [1, 4], Hybrid Search [1, 5] and Interpolation Search [1, 6] now-a-days which work only on sorted data. But till now there are less approaches that work on randomly

generated unsorted data. Several optimal data structures tools are badly required to operate this very large number of unsorted data which are producing day by day. Data Mining is one of the most important industrials topic now-a-days. Under this area of data mining for all social networks, governmental or non-governmental institutions and ecommerce industries produce tons of unsorted data and they are to utilize it. For utilizing this tons of unsorted data there is a need of some specific feature based unsorted data structures tools like searching algorithm [6] based on unsorted data array. So, a data structure tool [7] that will work on directly unsorted data is the prime concern for developing another searching approach.

Data scientists are trying to develop several data structures tools to utilize the tons of unsorted data randomly around the globe continuously. With a view to helping the data scientists here I am trying to develop data structures tools for finding out of any data from any given array of unsorted data. We know that sorting of data consumes a large number of time; so, from this concept of time utilization there needs some specific approach that can perform searching operation on unsorted data which minimize the uses of time. As the technology and technology related models/industries appreciate the approaches that minimize time consuming, so this is the demand of time to have high performer approach consuming less time without sorting the large data set at a single time.

Through this whole paper we will go through the approaches to develop a specific feature based searching algorithm entitled paw search algorithm that will be capable to perform searching operations on unsorted data, and here we may also go through the help of some existing searching approaches and sorting approaches at the inner phase of searching operation to ensure the high performance of searching.

The main principle of this Paw Search Algorithm is to work on (i) unsorted data segmenting the given array of data into several (ii) blocks.

Initially it starts working with  $x$  blocks of unsorted data by making the square root of the length  $n$  of the given array of data i.e.,  $x = \text{ceiling}[\sqrt{n}]$  where  $n$  is length of the given array of unsorted data.

This algorithm will never check all of the blocks of unsorted data linearly, rather than it will go the blocks of unsorted data all but as like as binary approach but not fully follow the binary approaches. And for the inner block operation we will also call here the merge sort approach for the better performance of this paw search algorithm.

## II. LITERATURE REVIEW

In this section we will go through the several existing searching algorithms, and most of them here work only on sorted data:

### A. Classification of Searching Techniques

There are several searching techniques present now-a-days. Depending on external and internal issue there are two types of searching techniques as (i) external search and (ii) internal search, and based on sequential and interval issue there are two types of searching techniques as (i) sequential search and (ii) interval search.

### B. Present Searching Algorithms

There are several searching algorithms based on sorted data. Some of them are listed below-

1) *Linear search algorithm*: Linear search algorithm [3] could be an easy search algorithmic program. It's a sequent search that performed on sequences of numbers that are ascending or down or unordered. And it checks every component of the whole list to look a specific information from the list. If the comparison is equal, then the search is stopped and declared productive. For a listing with  $n$  things, the most effective case is once the worth of item to be searched is adequate to the primary component of the list, during this case only one comparison is required. Worst case is once the worth isn't within the list or happens one time at the top of the list, during this case  $n$  comparisons are required.

2) *Binary search algorithm*: It is a quick search formula because the run-time quality is  $O(\log n)$ . Divide and conquer Principle is used here as its' search formula. This formula performs higher for sorted knowledge assortment. In binary search [8], we tend to 1st compare the key with the item within the middle position of the info assortment. If there's a match, we are able to come forthwith. If the secret's but middle key, then the item should lie the lower  $1/2$  the info collection; if it's bigger, then the item should lie the higher  $1/2$  the info assortment.

3) *Hybrid search algorithm*: Hybrid Search algorithmic [3, 9] rule combines properties of each linear search and binary search and provides a far better and economical algorithmic rule. This algorithmic rule may be accustomed search in associate degree unsorted array whereas taking less time as compared to the linear search algorithmic rule. As mentioned this algorithmic rule is combines 2 looking algorithms, viz. Linear Search and Binary Search. Like Hybrid Search algorithmic rule, the array is split into 2 sections so searched in every of the sections. The algorithmic rule starts with examination the key component to be searched with the 2 extreme components of the array, the primary and therefore

the last, further because the middle component. If a match is found, the index worth comes back. However, if it's not, the array is split into two sections, from the center index. Currently the search is meted out within the section on the left in a very similar method. The acute components and therefore the middle component of the left division are compared with the key worth for a match, that if found, returns the index worth. If not, the left section is once more divided into two components and this method goes on until a match is found within the left section. If no match is found within the left division, then the algorithmic rule moves on to the proper division, and therefore the same procedure is meted out to search out a match for the key worth. Now, if no worth is found that matches the key worth even when ransacking through all sections, then it's more divided and therefore the method repeats iteratively till it reaches the atomic state. If the worth isn't gift within the array, as a result of that the algorithmic rule returns -1.

4) *Interpolation search algorithm*: Interpolation search [2, 10] rule is improvement over Binary search. The binary search checks the part at middle index. However, interpolation search could search at completely different locations supported price of the search key. The weather should be in sorted order so as to implement interpolation search. As mentioned the Interpolation Search is Associate in Nursing improvement over Binary explore for instances, wherever the values in a very sorted array are uniformly distributed. Binary Search continuously goes to the center part to ascertain. On the opposite hand, interpolation search could head to completely different locations in line with the worth of the key being searched. For example, if the worth of the secret's nearer to the last part, interpolation search is probably going to start out search toward the tip facet.

5) *Jump search algorithm*: Jump search algorithmic [11, 12] rule, additionally known as block search algorithmic rule. Solely sorted list of array or table will use the Jump search algorithmic rule. In jump search algorithmic rule, it's not in any respect necessary to scan each component within the list as we have a tendency to liquidate linear search algorithmic rule. We have a tendency to simply check the  $m$  component and if it's but the key component, then we have a tendency to move to the  $m + m$  component, wherever all the components between the  $m$  and  $m + m$  component square measure skipped. This method is sustained till  $m$  component becomes adequate to or larger than key component known as boundary price. The worth of  $m$  is given by  $m = \sqrt{n}$ , wherever  $n$  is that the total range of components in associate array. Once the  $m$  components attain the boundary price, a linear search is finished to seek out the key price and its position within the array. And also the numbers of comparisons square measure adequate to  $(n/m + m - 1)$ . It should be noted that in Jump search algorithmic rule, a linear search is finished in reverse manner that's from boundary price to previous price of  $m$ .

Though there is a large number of searching approaches [13] on different aspects like strings [14], numeric values and

so on there is still a concern of optimizing [15, 16] these approaches.

Now-a-days industry requires specific feature based searching tools like audio, video and/or image based searching [17, 18], and as the industry is changing day by day with the help of upgraded technology, searching approaches are also gradually being changed as needed [19, 20].

And, still there needs of most powerful, high performer, fast searching unsorted data based searching approaches; keeping this conscious in mind, this paw search approach for unsorted data is demand of time now.

### III. METHODOLOGY

In this Methodology section, we will go through several sections like Planning, Design, Paw Search Algorithm etc. for the development of the proposed approach of search precisely and clearly:

#### A. Planning

Define To develop this algorithm I have planned several data structure approaches, arrays, sub-arrays or blocking, sorting approaches and so on.

- First plan is to manage several unsorted data sets that may be generated from different environment like weather data, space data and son.
- Second plan is to find out the length of the array with filling this array with that unsorted raw data.
- Third plan is to divide the unsorted array into several sub-arrays which are termed as data blocks in the later chapters of this paper.
- Fourth plan is to find out an optimal way to have operations on these blocks by traversing them.
- Fifth plan is to operate a searching approach on the blocks for finding out the optimal outputs.
- Sixth plan is to calculate the time and space complexity of this algorithm.
- Seventh plan is to compare these time and space complexity with different present searching algorithms properly.

The designation process of this algorithm is briefly described in part B of this section.

#### B. Design

To design this algorithm we are to go through a list of unsorted data set firstly as the main principle of this Paw Search - A Searching Approach for Unsorted Data Combining with Binary Search and Merge Sort Algorithm is to work on (i) unsorted data segmenting the given array of data into several (ii) blocks.

Initially it starts working with x blocks of unsorted data by making the square root of the length n of the given array of data i.e.,  $x = \text{ceiling}[\sqrt{n}]$  where n is length of the given array of unsorted data. but when the length of this array isn't a perfect square root number then the block number becomes a

fraction number, but the block number can't be a number as a fraction number in real, so we are to operate here the ceiling operator to get the integer number of blocks. But in this situation there needs some dummy data as like zero to make the block size perfect i.e., same length of each block.

For example let assume an unsorted array arr1[ ] of data of the length of 16 which is a perfect square root number that is shown in Table I.

TABLE I. UNSORTED ARRAY ARR[ ] OF 16 LENGTH

5	0	6	3	7	1	9	2	4	17	10	8	11	16	13	15
---	---	---	---	---	---	---	---	---	----	----	---	----	----	----	----

Here n=16; Since 16 is a perfect square root number

So the block numbers,  $x = \text{int}[\sqrt{16}] = 4$

TABLE II. X BLOCKS FROM ARR1[ ]

5	0	6	3	7	1	9	2	4	17	10	8	11	16	13	15
Block 1				Block 2				Block 3				Block 4			

Here the Block1, Block2, Block3 and Block4 are the four blocks of the segmented arr1 [ ] shown in Table II

Now let assume another unsorted array arr2 [ ] of data of the length of 8 which is not a perfect square root number that is shown in Table III

TABLE III. UNSORTED ARRAY ARR2[ ] OF 8 LENGTH

7	5	10	3	21	1	6	9
---	---	----	---	----	---	---	---

Here n=8; Since 8 is not a perfect square root number

So the block numbers,  $x = \text{int}[\sqrt{8}]$   
 $= \text{int}[2.828427125]$   
 $= 3$  [By applying ceiling operation]

TABLE IV. X BLOCKS FROM ARR2[ ]

7	5	10	3	21	1	6	9	0
Block 1			Block 2			Block 3		

Here the Block1, Block2 and Block3 are the three blocks of the segmented arr2[ ] shown in Table IV. In Block4 there is putted an extra zero as a dummy data for remaining the blocks size same.

However, this paw search algorithm will never visit all of the blocks of unsorted data linearly, rather than it will go through the blocks of unsorted data all but as like as binary approach. But it won't fully follow the binary approach.

The designing resources and working procedures list of this algorithm is listed here-

- Unsorted Data Set
- Square Root Generating Function
- Ceiling Operator

- Generating Blocks
- Block Visiting Loop
- Inner Block Searching Approach
- Exit

### C. Paw Search Algorithm

Assume that there is an array with the length of  $n$  of the nodes value of any given graph or other randomly generated unsorted data, now let's demonstrate our desired Paw Search Algorithm for finding out the target data from this given array of unsorted data. Here, we will go through the procedural steps of this Paw Search Algorithm. The procedures of this Paw Search Algorithm are shown below-

#### PAW (SEARCH ALGORITHM)

```

Divide the given array into x blocks where  $x = \text{ceiling}[\sqrt{n}]$ 
Loop for Block
//For Block1:
Block[ ] = x1[ ] = [ ]
    mergesort (Block[ ], L, U)
If(Block[last_element]  $\geq$  Target)
{
    If(x[last_element] == target)
        {Print "TARGET FOUND"}
        Exit
    Else
    {
        BinarySearch ( Block(x)[ ], 1,U)
            If (x[mid] == target)
                {Print "TARGET FOUND"}
                Exit
            Else
                {Jump to the next Block}
    }
Else
    {Jump Block[ ] = x[last]}
Update Loop
Loop Exit
If (target == not found)
    {Print "Unsuccessful"}
Exit
    
```

#### //Merge Sort Function

```

mergesort (Block[ ], l, U)
If U > 1
Find the middle point to divide the array into two halves:
middle m = 1+ (U-1)/2

    Call mergeSort for first half:
    Call mergeSort(arr, l, m)
    Call mergeSort for second half:
    Call mergeSort(arr, m+1, U)

Merge the two halves sorted in step 2 and 3:
Call merge(arr, l, m, U)
    
```

#### //Binary Search Function

```

Binary Search (Block[ ], 1,U)
Input the Block[ ] array of x elements I sorted form
LB = 0,UB = n; mid = int((LB+UB))/2
Repeat step 4 and 5 while(LB <= UB and (A[mid] != item)
If (item < A[mid]) UB = mid-1
Else
    LB = mid+1
    mid = int((LB+UB)/2)
    If (A[mid] == item)
        Print "Item is found"
    Else
        Print "Item is not found"
    Exit
    
```

The above mentioned procedures are the proposals of the Paw Search Algorithm, which also includes the Binary Search and Merge Sort Algorithm for completing its operation more efficiently. The further explanation of this algorithm is discussed later sections with proper examples.

### D. Explanation and Implementation

Let's understand the block visiting procedures now, a graphical view is illustrated in Fig. 1 to show the working flow of the  $x$  blocks generated from the length of the array by making square root on it, and the length of each block is also  $x$  i.e., the block size and the block numbers are same.

Here  $x$  is the block number and  $l, k, m, p, y$  are also the sub number of  $x$  and they are the right mid, right-right mid, ..... , left mid, left-right mid, ..... , and gradually so on.

For implementing this algorithm let assume an array  $A[]$  with the length of  $n$  as shown in the following Table V.

TABLE V. GIVEN ARRAY WITH N ELEMENTS

ARRAY ELEMENTS										----					
INDEX NUMBER	0	1	2	3	4	5	6	7	8	----	n-4	n-3	n-2	n-1	n

So the initial step of this algorithm is to calculate the square root value  $x$  of the length of the given array  $A[]$ .

$$x = \text{sqrt}(n) = \text{ceiling}(\sqrt{n})$$

It generates  $x$  number of blocks as Block(1), Block(2), Block(3), Block(4), ....., Block(x-3), Block(x-2), Block(x-1), Block(x) which are shown in Fig. 2.

The operational steps of these  $x$  number of blocks are also shown in Fig. 2. This algorithm follows the Left to Right approach. According to this Fig. 2, the first Block[1] is to go under the algorithmic operation firstly, secondly the last Block[x], then Block [mid], then Block [Right-mid], then Block [Right-mid], ....., ....., ....., Block [Left-mid], Block [Left-Right-mid], ....., Block [Left-Left-mid], ....., ....., ....., etc.

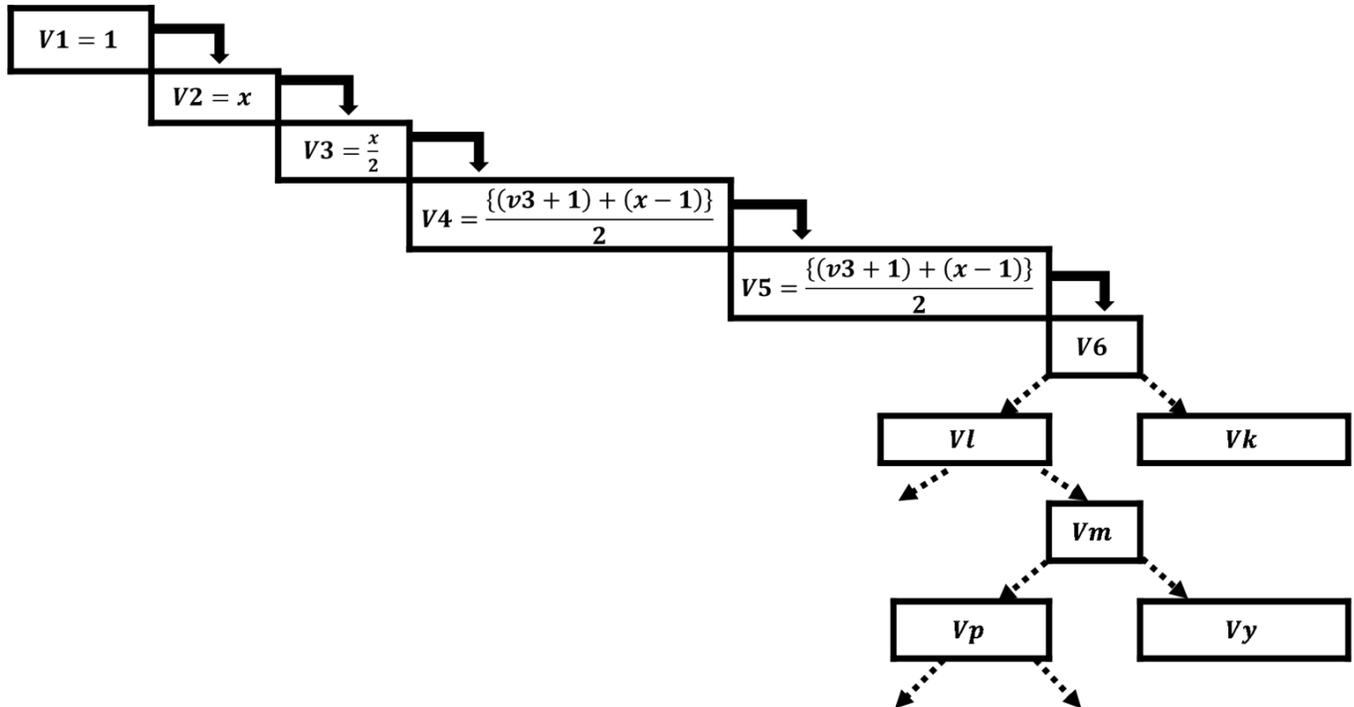


Fig. 1. Traversing procedure tree of the  $x$  blocks

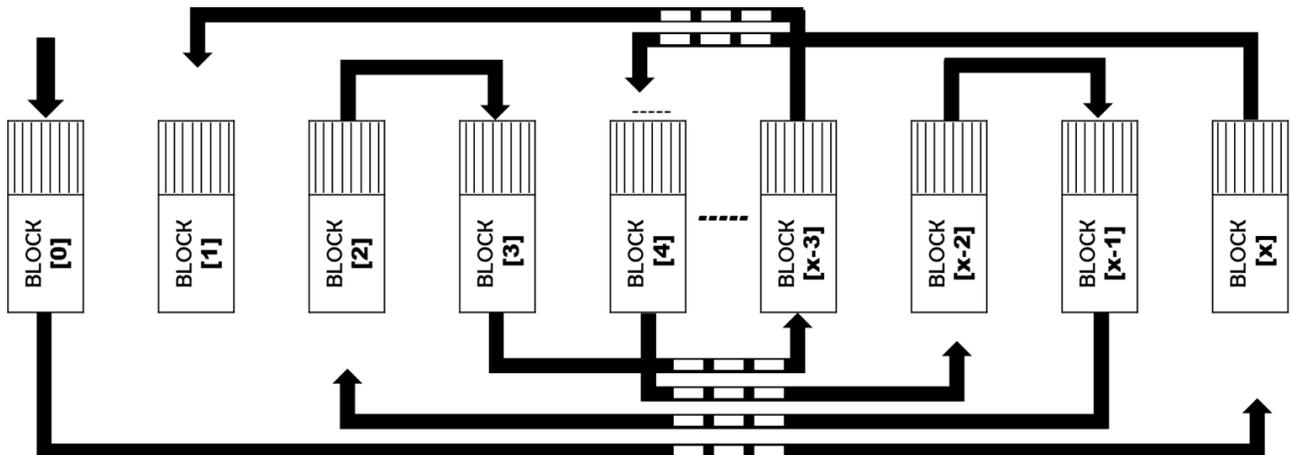


Fig. 2. Graphical view with paw search algorithm of the given array (table 5) with elements

For better understanding the implementation of this Paw Search Algorithm we will go through an example with proper explanation. Lets' assume another array B[] with the length of n, where n=9 as shown in Table VI. And let the SEARCH ITEM = Target = 5.

TABLE VI. ARRAY B[] WITH 9 ELEMENTS

4	6	1	9	3	5	8	7	2
---	---	---	---	---	---	---	---	---

Firstly, Lets' find out the number of Blocks:

$$x = \text{ceiling}[\sqrt{n}] = \text{ceiling}[\sqrt{9}] = 3$$

So, the blocks are shown in Table VII follows:

TABLE VII. BLOCKS OF ARRAY B[] WITH 9 ELEMENTS

4	6	1	9	3	5	8	7	2
Block 1			Block 2			Block 3		

For Block1:

```
//Sort the block using merge sort approach
mergesort (Block1[ ], 0, 2) //input
{
Block(x)[]={3, 5, 9} //sorted
}
If((Block1(last)==6) >= (Target==5))
{
Print "TARGET FOUND"
}
Else
{
//Binary Search
beg = lower_bound = 0
end = upper_bound = 2
mid =(beg + end)/2=(0 + 2)/2= 1
Block1[mid]= x[1]= 4
if (x[1]== target)
{
Print "TARGET FOUND"
Exit
}
Else
{
```

Jump to the next Block

```
}
}
Else
{
Jump x(last)
}
x++
Exit
```

For Block2:

```
//Sort the block using merge sort approach
mergesort (Block2[ ], 3, 5) //input
{
Block(x)[]={3, 5, 9} //sorted
}
If((Block1(last)==9) >= (Target==5))
{
If(x(last)==target)
{
Print "TARGET FOUND"
}
Else
{
//Binary Search
beg = lower_bound = 3
end = upper_bound = upper_bound+2=5
mid =(beg + end)/2=(3 + 2)/2= 4
Block1[mid]= x[4]= 5
if (x[4]== target)
{
Print "TARGET FOUND"
Exit
}
Else
{
Jump to the next Block
}
}
}
```

```

Else
{
Jump x(last)
}
x++
Exit
    
```

So, here the target is found in Block2.

#### IV. PERFORMANCE ANALYSIS

Some fundamental key terms related to performance measurement of this proposed searching approach of paw search algorithm will be discussed through this section briefly. Basically, here we will cover the time and space complexity of this proposed searching approach and also cover a brief comparison of different existing searching approaches with this proposed searching approach:

##### A. Space Complexity

The Now lets' go through the time complexity phase of this Paw Search Algorithm. For calculating Time complexity of this algorithm we are to go through the divide and conquer approach of recursive method through traversing the x blocks generated by squaring root the length n of the given array of data.

Let the block1 of the length of x elements generated by squaring root the length n of the given array of data i.e.,  
 $Block1(x) = \{E(1), E(2), E(3), E(4), \dots, E(x-3), E(x-2), E(x-1), E(x)\}.$

First of all we are to calculate the space complexity of merge sort approach for sorting this sub array i.e., Block1 of x length. And we already know that the space complexity of this merge sort approach is  $O(x)$  that means that it needs of space for sorting this sub array data is as equal as the length of the sub array, here which is x. As the size of each and every block is same and at a time only one block will be sorted, so here the space complexity is  $O(x)$ .

Now, let's calculate the space complexity of this paw search algorithm to find out the target value for this sub array x i.e., Block1

```

For Block1:
//Space Complexity Calculation
If(x(last)>=target)
{
If(x(last)==target)
{
Print "TARGET FOUND"
}
Else
{
    
```

```

Binary Search Algorithm
if (x[i]== target)
{
Print "TARGET FOUND"
i++
Exit
}
Else
{
Jump to the next Block
}
}
}
Else
{
Jump x(last)
}
Exit
    
```

So, there needs space as same as the length of the array x for performing this operation successfully. We can also see the graphical view (push and pop operation of stack method) of the recursive method of this Block1 x as below in Fig. 3.

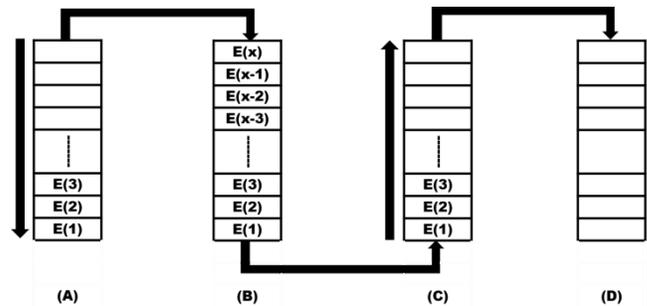


Fig. 3. Space complexity calculation

Where Fig. 3(A) shows the PUSHING of Block1's data into the STACK, Fig. 3(B) shows Block1 fully PUSHED into the STACK, Fig. 3(C) shows the POPPING of Block1's data from the STACK and Fig. 3(D) shows the Block1 which is fully popped from the STACK.

So, this Block1 needs same space as the length of this Block1 i.e., x. Similarly, all rest of the blocks need same space as their block size. So, here the space complexity is  $\log(x)$ .

Now, for finding out the target element from each block we will operate here the Binary Search approach. And we already know that the space complexity of the Binary Search approach is  $\log(x)$ .

So, the space complexity under the categories of worst case, average case and best case of this Paw Search Algorithm as below:

Paw\_(Space\_complexity)  $S(n) = O(\{\text{Space Complexity of the Block of given array} + \text{Space Complexity of Merge Sort Algorithm} + \text{Space Complexity of Binary Search Algorithm}\})$

$$\begin{aligned} \Rightarrow \text{Paw}_-(\text{Space\_complexity}) S(n) &= O(\log n + n + \log n) \\ &= O(n + 2 \log n) \\ &= O(2 \log n) \end{aligned}$$

$\therefore$  Paw\_(Space\_complexity)  $S(n) = O(\log n)$  [ $\because 2$  is the constant]

Hence the space complexity of this paw search algorithm is  $\log n$

### B. Time Complexity

Now let's go through the time complexity phase of this Paw Search Algorithm. For calculating Time complexity of this algorithm we are to go through the divide and conquer approach of recursive method through traversing the  $x$  blocks generated by squaring root the length  $n$  of the given array of data.

Let the Block1 of the length of  $x$  elements generated by squaring root the length  $n$  of the given array of data i.e.,  $\text{Block1}(x) = \{E(1), E(2), E(3), E(4), \dots, E(x-3), E(x-2), E(x-1), E(x)\}$ .

Firstly, we are to calculate an extra time for generating this  $x$  blocks by making square root of the given data array of the length of  $n$ .

Secondly, we are to consider the time complexity of merge sort approach for sorting this sub array i.e., Block1 of  $x$  length. And we already know that the time complexity of this merge sort approach is  $x \log(x)$ .

Now, let's calculate the space complexity of this paw search algorithm to find out the target value for this sub array  $x$  i.e., Block1

```
//For Block1:
If(x(last)>=target)
    If(x(last)==target)
        Print "TARGET FOUND"
    Exit
Else
    BinarySearch( Block[ ], L, U)
        if (x[i] == target)
            Print "TARGET FOUND"
        Exit
    Else
        Jump: Update Block
```

```
Else
    Jump: Update Block
Exit Block
Exit Loop
```

Thirdly, we are to find out the time complexity for the Binary Search approach for this Block1 i.e.,  $x$ . And we already know that the time complexity of the Binary Search approach is  $\log(x)$ .

So, the time complexity under the best case category is  $\sqrt{n}$

And the time complexity under the category of worst case and average case of this Paw Search Algorithm as bellow:

Paw\_(Time\_Complexity)  $T(n) = O(\{\text{Time for making Square Root of the given array} + \text{Time Complexity of Merge Sort Algorithm} + \text{Time Complexity of Binary Search Algorithm}\} * \text{Number of Blocks})$

$$\begin{aligned} \Rightarrow \text{Paw}_-(\text{Time\_Complexity}) T(n) &= O\{(1+n \log n + \log n) * \sqrt{n}\} \\ &= O\{\sqrt{n}(1+(n+1) \log n)\} \\ &= O\{\sqrt{n}(n \log n)\} \\ &= O\{n * \sqrt{n}(\log n)\} \\ &= O\{\sqrt{n^3}(\log n)\} \end{aligned}$$

$\therefore$  Paw\_(Time\_Complexity)  $T(n) = O(\sqrt{n^3} \log n)$

Hence the time complexity under the categories of worst case and average case of this paw search algorithm is  $\sqrt{n^3} \log n$

### C. Difference between Binary Search and Paw Search

The Paw Search Algorithm and the Binary Search Algorithm aren't same. There are several distinct difference between this two approaches. A difference chart between these two algorithms is shown in Table VIII follows:

TABLE VIII. DIFFERENCE BETWEEN PAW SEARCH AND BINARY SEARCH ALGORITHM

Paw Search Algorithm	Binary Search Algorithm
It begins its operation with the unsorted array of data.	It begins its operation with the sorted array of data.
It divides the given array of unsorted array of data of $n$ length into $x$ blocks by squaring root the length i.e., $x = \sqrt{n}$	It doesn't divide the array into blocks.
The input data is either unsorted or unsorted doesn't fact here.	The input data must be sorted here.
Time Complexity here in Worst Case is $\sqrt{n^3} \log n$	Time Complexity here in Worst Case is $\log n$
Time Complexity here in Average Case is $\sqrt{n^3} \log n$	Time Complexity here in Average Case is $\log n$
Time Complexity here in Best Case is $\sqrt{n}$	Time Complexity here in Best Case is $\log n$
Space Complexity here in Worst Case is $\log n$	Space Complexity here in Worst Case is $\log n$
Space Complexity here in Average Case is $\log n$	Space Complexity here in Average Case is $\log n$

Time Complexity here in Best Case is $\log n$	Time Complexity here in Best Case is $\log n$
It is a combined searching system.	It is a unique searching system.

So, the Paw and Binary Searching technique isn't similar at all rather than it is quite different and comparatively more efficient than Binary Searching technique. It also should be mentioned that the Paw Search Algorithm solves the limitation of taking fully sorted array as an input of Binary Search Algorithm.

**D. Difference between Jump Search and Paw Search**

The Paw Search Algorithm and the Jump Search Algorithm aren't same. There are several distinct difference between this two approaches. A difference chart between these two algorithms is shown in Table IX follows:

TABLE IX. DIFFERENCE BETWEEN PAW SEARCH AND JUMP SEARCH ALGORITHM

Paw Search Algorithm	Jump Search Algorithm
It begins its operation with the unsorted array of data.	It begins its operation with the sorted array of data.
It divides the given array of unsorted array of data of n length into x blocks by squaring root the length i.e., $x = \sqrt{rt}(n)$	It also divides the given array of sorted array of data of n length into x blocks by squaring root the length i.e., $x = \sqrt{rt}(n)$
It doesn't follow the linear approach for traversing its blocks.	It follows the linear approach for traversing its blocks.
It is faster.	It is comparatively slower.
It doesn't travel the blocks sequentially.	It travels the blocks sequentially.
Under the block operation it operates here binary search approach as an inner approach.	Under the block operation it operates here linear search approach as an inner approach.
The input data is either unsorted or unsorted doesn't fact here.	The input data must be sorted here.
It is a combined searching system.	It is a unique searching system.

So, the Paw and Jump Searching technique isn't similar at all rather than it is quite different and comparatively more efficient than Jump Searching technique. It also should be mentioned that the Paw Search Algorithm solves the limitation of taking fully sorted array as an input of Jump Search Algorithm.

**E. Time Complexity Comparisons with others Algorithms**

A comparison list of time complexity of different search algorithms like linear search, binary search, hybrid search, interpolation search and paw search in different cases like worst case, average case and best case is shown in Table X as follows:

TABLE X. TIME COMPLEXITY COMPARISONS

	Best Case	Average Case	Worst Case
Linear Search	$O(1)$	$O(n)$	$O(n)$
Binary Search	$O(1)$	$O(\log n)$	$O(\log n)$
Hybrid Search	$O(1)$	$O(\log 2n)$	$O(n)$

Interpolation Search	$O(1)$	$O(\log(\log N))$	$O(n)$

**F. Space Complexity Comparisons with Others Algorithms**

A comparison list of space complexity of different search algorithms like linear search, binary search, hybrid search, interpolation search and paw search in different cases like worst case, average case and best case is shown in Table XI as follows:

TABLE XI. SPACE COMPLEXITY COMPARISONS

	Best Case	Average Case	Worst Case
Linear Search	$O(1)$	$O(n)$	$O(n)$
Binary Search	$O(1)$	$O(\log n)$	$O(\log n)$
Hybrid Search	$O(1)$	$O(\log 2n)$	$O(n)$
Interpolation Search	$O(1)$	$O(\log(\log N))$	$O(n)$
Paw Search	$O(\log n)$	$O(\log n)$	$O(\log n)$

**V. CONCLUSION WITH FUTURE WORK**

By developing this long discussion of this research paper, I come to know that research is the fundamental weapon of this globalizing world i.e., IT world, and the large number of unsorted data is the heart of each and every research now-a-days. And managing this large number of unsorted data properly with proper searching technique is the core point of this paw search algorithm. The prime attraction of this research work is to develop a specific as well as more optimal formula of searching purposes from the unsorted list or array of data with the help of other searching and sorting techniques like merge sort and binary search. This Paw Search Algorithm shows the optimal way to generate a proper searching output taking an unsorted data list or array of data along with optimal time and space complexity, several comparisons of different searching approaches with this paw search algorithm are shown in Table VIII, IX, X and XI consecutively.

However, research is a continuous process. It will be upgraded with the demand of time day by day. There are also available a lot of future works here, some of them are listed below:

- Developing a more optimal logic/formula to optimize this algorithm
- Developing a Machine Learning Model to predict the desired block containing the desired data with Machine Learning Approach

**ACKNOWLEDGMENT**

It is a great pleasure for me to present this thesis paper titled as "Paw Search - A Searching Approach for Unsorted Data Combining with Binary Search and Merge Sort Algorithm".

I express heartiest thanks to friends and my well-wisher for their continuous inspiration and support, which led me to complete this research work.

Finally, I express my appreciation to my parents and other family members for their unconditional support as without their support and inspiration, it would be impossible for me to complete this research successfully.

#### REFERENCES

- [1] Sultana, N., Paira, S., Chandra, S., & Alam, S. S. (2017, February). A brief study and analysis of different searching algorithms. In 2017 Second International Conference on Electrical, Computer and Communication Technologies (ICECCT) (pp. 1-4). IEEE.
- [2] Roopa, K., & Reshma, J. (2018). A comparative study of sorting and searching algorithms. International Research Journal of Engineering and Technology (IRJET).
- [3] Das, P., & Khilar, P. M. (2013). A randomized searching algorithm and its performance analysis with binary search and linear search algorithms. International Journal of Computer Science & Applications (TIJCSA), 1(11).
- [4] Pathak, A. (2015). Analysis and Comparative Study of Searching Techniques. International Journal of Engineering Sciences & Research Technology, 4(3), 235-237.
- [5] Subbarayudu, B., Gayatri, L. L., Nidhi, P. S., Ramesh, P., Reddy, R. G., & Reddy, C. K. K. (2017). Comparative analysis on sorting and searching algorithms. International Journal of Civil Engineering and Technology (IJCIET), 8(8), 955-978.
- [6] Rahim, R., Nurarif, S., Ramadhan, M., Aisyah, S., & Purba, W. (2017, December). Comparison searching process of linear, binary and interpolation algorithm. In Journal of Physics: Conference Series (Vol. 930, No. 1, p. 012007). IOP Publishing.
- [7] Data Structures, Seymour Lipschutz and G A Vijayalakshmi Pai, SCHAUMS'S OUTLINES, 2013-2014.
- [8] Jacob, A. E., Ashodariya, N., & Dhongade, A. (2017, August). Hybrid search algorithm: Combined linear and binary search algorithm. In 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS) (pp. 1543-1547). IEEE.
- [9] Introduction to Algorithms, Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, Clifford Stein, Third Edition, 2017-2018
- [10] Harman, M., & McMinn, P. (2009). A theoretical and empirical study of search-based testing: Local, global, and hybrid search. IEEE Transactions on Software Engineering, 36(2), 226-247.
- [11] Shneiderman, B. (1978). Jump searching: A fast sequential search technique. Communications of the ACM, 21(10), 831-834.
- [12] Mahboob, T., Akhtar, F., Asif, M., Siddique, N., & Sikandar, B. (2015). Survey and Analysis of Searching Algorithms. International Journal of Computer Science Issues (IJCSI), 12(3), 169.
- [13] Boyer, R. S., & Moore, J. S. (1977). A fast string searching algorithm. Communications of the ACM, 20(10), 762-772.
- [14] Bentley, J. L., & Sedgewick, R. (1997, January). Fast algorithms for sorting and searching strings. In Proceedings of the eighth annual ACM-SIAM symposium on Discrete algorithms (pp. 360-369).
- [15] Mehlhorn, K., Sanders, P., & Sanders, P. (2008). Algorithms and data structures: The basic toolbox (Vol. 55, p. 56). Berlin: Springer.
- [16] Tuparov, G., Tuparova, D., & Jordanov, V. (2014). Teaching sorting and searching algorithms through simulation-based learning objects in an introductory programming course. Procedia-Social and Behavioral Sciences, 116, 2962-2966.
- [17] Wang, A. (2003, October). An industrial strength audio search algorithm. In Ismir (Vol. 2003, pp. 7-13).
- [18] Zabinsky, Z. B. (2009). Random search algorithms. Department of Industrial and Systems Engineering, University of Washington, USA.
- [19] Shareef, H., Ibrahim, A. A., & Mutlag, A. H. (2015). Lightning search algorithm. Applied Soft Computing, 36, 315-333.
- [20] Bentley, J. L., & Yao, A. C. C. (1976). An almost optimal algorithm for unbounded searching. Information processing letters, 5(SLAC-PUB-1679).

# A Survey of Forensic Analysis and Information Visualization Approach for Instant Messaging Applications

Shahnaz Pirzada<sup>1</sup>, Nurul Hidayah Ab Rahman<sup>2</sup>, Niken Dwi Wahyu Cahyani<sup>3</sup>, Muhammad Fakri Othman<sup>4</sup>  
Centre for Information Security Research, Universiti Tun Hussein Onn Malaysia, Parit Raja, Malaysia<sup>1,2</sup>  
School of Informatics, Telkom University, Bandung, Indonesia<sup>3</sup>  
Application & Research on Multimedia, Universiti Tun Hussein Onn Malaysia, Parit Raja, Malaysia<sup>4</sup>

**Abstract**—Instant messaging applications, including WhatsApp, Viber, and WeChat, are moving beyond text messages to videos and voice calls, which are proportioned to current media, files, and locations. In this study, we surveyed existing forensic visualization and forensic analysis techniques for instant messaging applications, with the aim of contributing to the knowledge in the discussion of these research issues. A total of 61 publications were reviewed after searching various academic databases, including the IEEE, ACM Digital Library, Google Scholar, and Science Direct during the last five years. Our observation from research trends indicates that both forensic analysis and information visualization are relatively mature research areas. However, there is a growing interest in forensic visualization and automated IM forensic analysis. We also identified the lack of discussion on forensic selection criteria in existing forensic visualization works and the needs of benchmarking the evaluation method of automate forensic analysis tools.

**Keywords**—Forensic analysis; forensic visualization; instant messaging apps; mobile forensics; and mobile communication apps

## I. INTRODUCTION

Mobile Instant Messaging (IM) applications (apps) are becoming essential for smartphone users in their daily communication activities. As reported by Statista [1], the number of smartphone subscriptions worldwide in 2021 surpassed more than seven billion. Some of the most widely used IM apps include LINE, WhatsApp, WeChat, and Facebook Messenger [2]. As an example, WhatsApp has been upgraded beyond a basic messaging app to support more sophisticated features such as end-to-end encryption, deleting sent messages, and enable disappearing messages. These features, however, could be exploited by cybercriminals targeting IM apps for criminal activities.

According to a Norton report, malware, keylogging, and social engineering are the top three potential cybersecurity risks related to IM apps [1]. This is consistent with a report by Kaspersky, which showed that 341,954 attempts to follow phishing links were blocked in 2021, with 90% links coming from WhatsApp [2]. Furthermore, phishing statistics reported by PurpleSec identified that WhatsApp is one of the top three most impersonated brands in phishing attacks [3], [4].

In the context of smartphones, the acquisition of digital evidence involves mobile forensic techniques [7]. The acquired

artifacts can be valuable, as these include various significant metadata, such as application data, communication data, location data, and browsing history data. These data are, however, produced in unstructured data - raw data that are not in the organized data model form. The unstructured data could be challenging for forensic examination activities, for instance time-consuming and increase investigation cost [8], [9].

Therefore, there has been significant interest in examining appropriate approaches to expedite digital forensic analysis activities. One of the potential approaches is the use of forensic visualization. It is a common practice for digital forensic investigators to perform cross-analysis using various forensic software, but there is a lack of advanced visualization approaches to facilitate evidence analysis [5]. It also has been pointed out that the application of multimedia technology in presenting digital evidence could increase judicial understanding [6], [7], [8].

In this study, we explored the literature on forensic analysis of mobile applications, information visualization, and forensic visualization. The contributions of this work are twofold: (1) to provide insights into digital forensic analysis and the development of its automated tools, forensic analysis of IM apps, and forensic visualization to assist forensic analysis, and (2) to discuss the research trends and future research directions for these areas, including the potential of incorporating forensic visualization in IM forensic analysis. The knowledge gaps are identified from this study such as the needs of evaluation benchmarking and the lack of forensic selection criteria in the existing studies.

The remaining sections of this survey paper are organized as follows: Section II presents the review methodology while in Section III, associated works addressing forensic analysis and the techniques of visualization and mobile forensic analysis are discussed. Section IV reviews existing work in forensic analysis of IM apps. Section V discusses the role of forensic visualization in forensic analysis. Section VI presents the discussion on the research trends of forensics visualization and mobile forensic analysis techniques for future works. Section VII concludes this study.

## II. REVIEW METHODOLOGY AND PROCESS

A literature survey in forensic analysis, forensic visualization, and information visualization was performed by adopting the method used in [9] and [10] (see Fig. 1).

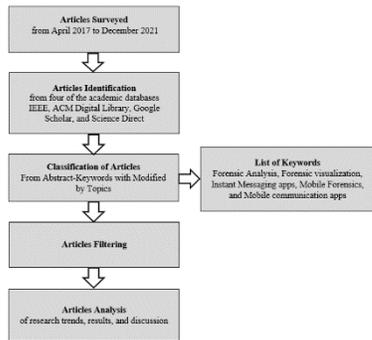


Fig. 1. Data collection method

To obtain a fair overview of the literature on forensic analysis and information visualization, we surveyed materials published in the English language over the past five years (i.e., April 2017 to June 2022). A total of 61 publications were located after searching various academic databases, including the IEEE, ACM Digital Library, Google Scholar, and Science Direct.

The search words used were different in each case, such as “digital forensics”, “forensic analysis”, “intelligent visual analytics digital evidence”, “mobile forensics”, and “information visualization”. For the search in Title, Abstract, and Keywords, quotation marks were entered and modified by topic, such as “forensics visualization”, “information visualization in digital forensics”, “forensic analysis in visualization”, and “forensic analysis in mobile application”. We filtered the articles from the search results to include “digital forensics”, “forensic analysis”, and “visual analytics”. We utilized the term “Web of Science” in title, abstract, and keyword searches in Advanced Search, and searched within Topics. Furthermore, we defined the document type as “articles” with no restrictions for all search results, and we only looked at journal publications.

## III. DIGITAL FORENSICS ANALYSIS

Forensic analysis is one of the phases in digital forensics, which is undertaken after evidence collection and examination. It involves “analyzing the results of the examination, using legally justifiable methods and techniques, to derive useful information that addresses the questions that were the impetus for performing the collection and examination.” [11]. More importantly, forensic analysis encompasses the gathering of evidentiary materials, evidence interpretation, results validation, and evidence presentation in an intelligible manner [12].

As discussed in [13], event reconstruction includes the combination of temporal, functional, and relational analyses of available evidence artifacts. Temporal analysis includes searching for other events that occurred around similar timestamps as those of one or more events already identified as related to the case being investigated. Functional analysis

involves understanding what actions were possible within the environment of the offense and how the offender’s toolkit works. Relational analysis involves studying how the various systems involved in a compromise relate to each other and how they interact.

String and keyword search, file filtering, and timeline analysis are examples of commonly used techniques in forensic analysis. String and keyword searching techniques can be used to filter out words, phrases, strings, and keywords that provide clues when searching for evidence. It is one of the primary features used in both commercial and non-commercial forensic tools, (e.g. Magnet Axiom, Autopsy). It has also been widely validated in academic forensic investigation studies, such as those analyzing web URL information [14]. Nowadays, the conventional keyword searching technique might be limited in a large volume of data, as it could lead to false negative or false positive and requires background knowledge about the case [15]. Therefore, studies such as [16] and [17], have examined the use of the semantic-based approach for text clustering with the aim of improving the performance and accuracy of forensic analysis.

Another technique, known as file filtering, can be applied as digital sieving of important files from irrelevant files by utilizing cryptographic hashes to screen the hash values of target files. For example, when using the MD5 Hash (and/or SHA-1 Hash) and Known File Filter options to process evidence, a hash value for each item is generated. The hash value of each file item inside the evidence is computed, and known files are filtered for the freshly computed hash value data [18]. The file filtering technique is significant in forensically examining file systems, for example, examining file similarities, as in [19], and examining file types, as in [20]. A limitation of this technique, however, is that it cannot be applied to corrupted files.

The timeline analysis technique is a chronological analysis of incidents to display all occurrences in a chronological sequence. However, emerging computing pose various factors that must be considered to generate a unified timeline, such as different time zone settings, timestamp interpretations, clock skew, and syntax [21]. Therefore, manual timestamp inspections may no longer be adequate to support investigations. Some recent studies have applied other scientific methods to enhance temporal analysis such as visualization approach, such as using graph-based and ontology-based approaches [22], [23] and highlighting patterns in the timeline analysis [24], [25], [26].

It is unlikely that a single method of data analysis is applied when examining digital evidence, as the evidence could have multiple interactions. The applied methods of data analysis can be mixed, depending on the complexity of a system’s architecture. For example, Carrier [27] presented data analysis based on layers, in which physical storage media analysis was the bottom layer. The next layers comprised volume analysis, memory analysis, file system analysis, and database analysis, while application analysis was the top layer. In a work by [28], the authors demonstrated system analysis, operating system analysis, application analysis, network analysis, device analysis, and Shim cache analysis to present Windows registry

forensics. Analyzing cloud apps on Android mobile devices, [29] demonstrated the analysis of app files in private storage, external storage, app database, and app account data and analyzed apps using the static and dynamic analysis approach. X. Zhang et al. [30] performed IoT botnet forensics by applying network traffic analysis, servers' disk analysis, servers' memory analysis, and database analysis. As described in these works, the configurations and deployments of systems architectures can vary. Undertaking the tasks manually may result in advertent or inadvertent mistakes and biases. The advancement of computing infrastructures and the interconnection of devices have made the tasks more complex and time-consuming. Hence, the automation of the tasks is essential to improve the efficiency of forensic analysis.

**B. Metrics of Automated Forensic Analysis Tools**

The development of automated forensic analysis tools is a developing field with a wide range of scientific techniques applied in many cyber forensics areas to keep pace with evolving computer generations [31]. Ayers [32] proposed seven metrics for computer forensic tools that are absolute speed, relative speed, reliability, accuracy, completeness, auditability, and repeatability.

Absolute speed, which refers to the elapsed time required to complete analysis tasks. Relative speed, which involves the average rate at which the tool can process evidence compared with the rate at which data can be read from the original evidential media. Reliability, which includes the proportion of tests that the tool executes successfully, as in performing without crashes and providing outputs in the intended format. Accuracy, which refers to the proportion of analysis results that are correct. Completeness, which concerns the proportion of evidence artifacts present in a forensic image that are identified and reported by the tool. Auditability, which includes the proportion of results that are fully auditable back to the original evidence data. Repeatability, which involves the proportion of tests that ran as stipulated in every aspect. Table I summarizes the applied forensic tools metrics in existing works related to forensic analysis tools.

TABLE I. A SUMMARY OF FORENSIC TOOL METRICS USED

Metric	A. Singh et al. [28]	Lin et al. [33]	Kumar et al. [34]	Subedi et al. [35]	Anglano et al. [36]
Absolute speed	√	√ (efficiency)	-	-	-
Relative speed	-	-	-	-	-
Reliability	√	-	-	-	-
Accuracy	-	√	√	√	√
Completeness	√	√	-	-	√
Auditability	-	-	-	-	-
Repeatability	√	-	-	-	√ (fidelity)
Other metrics	-	effective	-		effectiveness, generality, repeatability

In practice, it is unlikely for a study to include all the specified metrics in a tool due to the limited scope, time, and tool functionality. It has been observed that completeness and accuracy are the most applied metrics. The metrics are consistent with digital evidence principles by the RFC 3227 [37].

There are also other studies that proposed new or additional metrics to evaluate forensic tools. For example, Lin et al. [33] used the effective metric to evaluate their tool's efficacy in locating the source of the evidence artifact. Additional metrics proposed by Anglano et al. [36] are (1) effectiveness, which is the ability to correlate users' actions with the generated data, and (2) generality, which is the ability to analyze any mobile application on as many different Android devices as possible. It is observed that this work defined the repeatability metric as the ability to provide to a third party the possibility of replicating the same set of experiments. Considering the repeatability definition by Anglano et al. [36], we argue that all works shown in Table I are repeatable because detailed methodologies were provided. Furthermore, it is essential for academic work to provide a repeatable methodology for comparison with other similar research works. Other notable observations are that: (1) aforementioned studies used the same term to discuss different evaluations, such as Anglano et al. [36] using the "effective" term to discuss correlation ability, while Lin et al. [33] used the term to discuss the ability to locate data, and (2) the studies used different terms to describe the same evaluation, for example, Anglano et al. [36] using "fidelity" as a term to discuss repeatability.

**IV. FORENSIC ANALYSIS IN MOBILE COMMUNICATION APPS**

Examining IM apps is one of the continuous research works in mobile forensics. In addition, new updates of apps' features pose challenges for mobile forensics practices [38]. In this section, we describe the basic concepts of mobile forensics, and recent works involving the forensic analysis of mobile IM apps are reviewed.

**A. Forensic Analysis of Mobile Instant Messaging Apps**

Forensic analysis of mobile IM apps has received considerable attention and is rapidly emerging in recent years. For example, Anglano et al. [39] performed an in-depth analysis on how to decode, interpret, and correlate the data generated by users, using Telegram Messenger as a case study. The experiments involved user account information, contacts, chats, message exchanges, phone calls, and deleted information. The proposed methodology of this work was evaluated based on completeness, repeatability, and generality. It should be noted that analysis tasks were conducted manually and involved the evaluation of sources. The study's event reconstruction includes temporal and relational analyses. In a similar vent, Akinbi and Ojie [40] demonstrated forensic analysis of Conversation and Xabber apps on Android smartphones. The study discussed the results obtained from the completeness of recovered data, the sources of data, and timeline analysis to present the chronology of chat logs, message contents, and deleted files.

Riadi and Firdonsyah [41] compared mobile forensics tools' effectiveness on four IM apps, which were Short Message Service, Blackberry Messenger (BBM), LINE, and WhatsApp. The applied tools in this study were Andriller, Oxygen Forensic Suite, WhatsApp DB/Key Extractor, and Metasploit. The experiments were conducted on Android smartphones and an Android smartwatch. Tool effectiveness was evaluated using the success rate percentage of artifact extraction, in which Oxygen Forensic Suite showed the highest success rate at 57.14% on BBM and WhatsApp, while Smartwatch achieved 42.85% success rate on SMS and LINE. Additionally, the completeness of the recovered artifacts was evaluated using the percentage of artifact extraction.

Due to concerns over privacy, the features of encrypted chat and end-to-end encryption have been updated in some IM apps. These pose another dimension in forensic analysis of investigating possible ways to recover data from encrypted databases. For example, [42] examined the encryption status of WhatsApp, Facebook Messenger, LINE, and Hangouts on Android smartphones. The study compared encrypted and unencrypted databases between rooted and unrooted smartphones. The results of the recovered data from unencrypted databases were presented in terms of the point of origin of data, database structures, and data completeness. Rathi et al. [43] performed forensic analysis tasks on four encrypted Android IM applications, which were WeChat, Telegram, Viber, and WhatsApp. It was demonstrated that encrypted WeChat and Viber databases can be acquired from rooted devices, WeChat database can be decrypted using the IMEI number and the phone identifier encryption key, WhatsApp messages can be acquired from unrooted devices, WeChat data can possibly be retrieved from unrooted devices by downgrading the app version, and Telegram data were irretrievable from unrooted devices. However, with the current version of the WeChat application, the approaches used in their study are no longer valid [44]. Although this study did not directly specify the evaluation method, it has been observed that the forensic analysis result was presented based on the completeness criterion. The analysis tasks were undertaken manually and involved the evaluation of evidence sources.

Also focusing on the decryption methodology for forensic analysis, [45] investigated Wickr and Private Text Messaging apps on both rooted and unrooted Android smartphones, as well as jailbroken and non-jailbroken iOS smartphones. The results of this study showed that the proposed method was able to decrypt the databases of both apps and examine the encrypted databases' structure. The verification method of recovered user-entered password was used to decrypt the entire data. The method was evaluated based on the efficiency of estimated password recovery time. It is observed that this work focused on cryptographic contribution, as there was no discussion on data completeness and sources of evidence.

There are studies that attempted to compare recovered data from volatile and non-volatile memories. Focusing on Android LINE Messenger, [46] examined data remnant through the simulation of user activities, such as installation, uninstallation, logins, conversations, file transfer, and other LINE activities. The findings of the study showed that evidence artifacts from the LINE app can be recovered from both volatile and non-

volatile memories. Similarly, their study attempted to analyze the artifacts using the evaluation metrics of evidence sources and completeness. Agrawal and Tapaswi [47] conducted the forensic analysis of the Android Google Allo messaging app. The study demonstrated manual analysis tasks on device images. An interesting observation is that this study applied inferential statistics to evaluate the completeness metric. The recovered files were evaluated based on the point estimation value, while margin of error was used to calculate confidence limits. Another study that applied the inferential statistics concept was conducted in [48]. The study demonstrated the residual data of Android Kik Messenger on the NAND flash memory and the heap memory. The results of this study were presented by comparing the count and the average number of recovered messages between the NAND memory and the RAM memory.

Considering the extensive manual tasks of forensic analysis, some studies have proposed the development of automated tools. Barradas et al. [49] demonstrated forensic analysis tasks on eight messaging apps, which were Facebook Messenger, WhatsApp, Viber, Signal, Twitter, Telegram, Hangouts, and Trillian. The proposed tool, called RAM Analysis System (RAMAS), was designed based on the file-carving approach to extract potential evidence artifacts from physical memory. This study simulated common user activities on messaging apps. The performance of RAMAS was measured using the analysis time metric by varying the memory image size and the number of modules. The results indicated that reducing the set of strings improved the elapsed time, while running several modules in parallel resulted in a sub-linear time for analysis completion. The tool also supported event reconstruction tasks by applying the timeline analysis technique.

Nizam et al. [50] presented a tool to automate keyword indexing to assist the forensic analysis of WhatsApp. The main module of the tool involved loading a list of keywords from selected crime categories and uploading WhatsApp chat text files. Subsequently, a text-matching algorithm was executed, and a keyword count that showed the number of found keywords in the chat files was generated. The tool was tested using software engineering properties, which were application functionality testing and user acceptance testing. No event reconstruction feature was observed in this tool.

Using WhatsApp and LINE as case studies, [51] proposed a tool that applied the information visualization approach to support forensic analysis activities. Using visualization techniques, the tool was able to visualize key information, such as total number of messages, number of contacts, top-most frequent contact, top-most frequent words, and location map. Event reconstruction was supported by this tool using the timeline analysis approach. As in [50], this tool was evaluated through functionality testing and user acceptance testing.

A notable observation found from these works is that the simulation of user activities on the studied apps was used to elicit the generation of artifacts. The recovered artifacts from IM apps were mainly associated with user account information, contact information, chat histories, message exchanges, media exchanges, phone calls, app database, and deleted information.

Furthermore, analysis results from rooted and unrooted Android devices or jailbroken and non-jailbroken iOS devices presented significant differences involving encrypted data. Most of the studies examined recovered artifacts on device images manually, except for a few studies that used proposed automated tools. We also note that results from the metrics of origin of artifact and completeness of recovered artifact were the two predominantly reported forensic analysis results. Evaluation metrics, however, varied in these studies, and several studies did not directly specify their metrics. We discuss our further observation of the trends of studies on forensic analysis of IM apps in Section V.

## V. AN OVERVIEW OF INFORMATION VISUALIZATION

Information visualization is the process of displaying data as graphical markings on a computer screen or other media, to enhance people's ability to recognize visual patterns, such as watching, browsing, discriminating, and comprehending data [33]. Information visualization (InfoVis) is considered as a mature area, as evidenced in a study by Rees and Laramee [52], which surveyed over 23,000 pages of information visualization books. The study also highlighted that InfoVis research papers were cited in many areas, including networks, finance, healthcare, and security.

Data types in InfoVis can be classified into multi/high-dimensional, relational, sequential/temporal, geospatial, and textual [53]. Multi-dimensional and high-dimensional data are presented in a table-like form, where the rows denote data objects, while the columns denote data dimensions, attributes, features, or descriptors. Relational data refers to the common case of binary relations and are represented as graphs. Sequential and temporal data concern the serial order of data points in a sequence, for instance, the time series data. Geospatial data involves creating maps of the real world to visualize spatial and non-spatial relationships among the data. Textual data are inherently multivariate data sources, for example, text corpora as a semi-structured source of information integrated with approaches (e.g., semantic, text-mining) to transform raw texts into structured data sources.

Sorapure [54] discussed four key elements in InfoVis for improving data interpretation, which are text, image, data, and interaction. Text is often included as titles, labels, annotations, explanations, and other commentaries. Significant functions of text are guiding interpretation, providing explanation, establishing context, and facilitating navigation. The image element concerns generating and maintaining users' interest by selecting the most effective ways to convey information. Data literacy is important in enabling users to formulate questions and make decisions informed by the data. It involves creating arguments based on the data, effectively using tools to manipulate and represent the data, and being able to communicate with the data. Interactions include activities from users, such as selecting, exploring, reconfiguring, encoding, abstracting, filtering, and connecting, to establish interactions with InfoVis.

The application of InfoVis in many different fields further indicates that it is a reliable approach to support data interpretation and decision-making tasks. In the context of digital forensic analysis, interpreting gathered evidence is the

key to establishing connections in crime investigations. With the use of intelligent computing and decision-support frameworks, InfoVis has a significant role in assisting forensic analysts in digital evidence analysis activities [55].

### A. Role of Visualization for Forensic Analysis

The use of information visualization techniques in digital forensics has received considerable interest in the digital forensics' community. Existing works applied information visualization to visualize forensic data using various case studies, such as mobile phones, network data, IoT data, and Windows system files.

A study in [56] demonstrated how visual representation could support faster and more accurate decision-making during real-time digital forensics investigations. The Nested Blocks and Guidelines Model (NGBM) was adopted to design the visualization interactions in this study. Using fileless malware as a case study, the usability of the proposed tool was evaluated through the tool's components, such as investigation timeline, network activity, read/write entropy, and system performance. Examples of the used data were "time series", "IP Address", "Windows Event Logs", "PowerShell Events", and "Syslog", while the involved visualization methods were line charts, ellipses, area charts, and time series.

Tassone et al. [5] proposed a proof of concept to visualize digital forensics datasets that consisted of three stages of visualization lifecycle, which were decode, store, and visualize. Three visualization techniques were used, namely treemap, geographic map plot, and word cloud. This study used three case studies to represent the XRY mobile forensics dataset to evaluate tool utility. The treemap technique was used to visualize SMS messages, the geographic map technique was used to map locations from a coupon app, and the word cloud technique was applied to visualize data from a text-based communication app. The three case studies were able to demonstrate the utility of the tool.

Kotenko et al. [57] proposed a visual analytics approach for network forensics to analyze network traffic. The proposed approach consisted of two stages, which are data slice classification and the selection of an information visualization model. The visualization model can be determined from data types, such as numerical (e.g., pie chart, bar chart), tree (e.g., TreeMaps), planar (e.g., Voronoi Maps), semi-structured (e.g., Chord diagram), unstructured (e.g., graph), and a combination of data types and models. The usability of the proposed tool was evaluated using a case study of an SSL-strip attack that involved 200584 network packets in the files. The model was demonstrated in a three-hour training lab session, in which the result showed that 8 out of 10 students were able to solve the use-case within the lab session using the proposed visual analysis tool.

Also investigating network traffic data, [58] applied a 3D model and the use of time-based information as a display third axis, combined with a computer network topology in a single interactive data. The proposed tool, Scanmap3D, was evaluated based on its effectiveness to solve questions from network forensics challenges. The results were compared with other

tools that applied traditional statistical and 2D graphical analysis approaches.

Implementing Windows Jump List, [59] presented a graphical digital forensics tool, known as Jump List Analyzer. The study applied statistical charts (e.g., histogram) to visualize attributes, such as AppID, time, zoom, CustDest file, recorded file, and GUI interface. The tool was demonstrated to support forensic investigations of users' background and behavior analysis. Compatibility, friendliness, and functionality were the three applied metrics to evaluate the Jump List Analyzer. A comparison with other tools suggested that the proposed tool can effectively visualize large volumes of data.

Analyzing an online social network, [60] proposed a forensic analysis model that included evidence acquisition, evidence solidification, evidence analysis, and evidence visualization. Evidence analysis used semantic analysis method in natural language processing, and it involved text analysis, hot word frequency analysis, and physical locations. Sina Microblog was used to demonstrate the proposed tool's feasibility. Web page files, such as HTML text data, CSS files, JavaScript files, and images, were the involved metadata. However, there was no detailed information on InfoVis.

Applying the use of 3D and 2D models, [61] proposed a drone forensic framework to investigate the post-flight investigation of drone activities. The authors applied 3D visualization models to visualize three specific parameters, which were roll, pitch, and yaw along the flight path, while other parameters, which were drone-controller communication, signal type, battery, altitude, number of satellites used, and speed at each point of time, were visualized using 2D models. Each parameter value was logged based on the timestamp. Graphs and charts, such as line charts, were utilized for the 2D models. This study measured forensic visualization aspects based on performance and responsiveness. The study indicated that the tool could manage the visualizations of sensor data without interruptions.

Focusing on cloud computing's containers and Virtual Machines (VMs), [62] applied InfoVis to visualize the bytes contained in a virtual machine file for rapid incident response. The extracted data were visualized using two-dimensional colored visualization. The proposed visualization method was evaluated based on relative speed and accuracy using a series of *t*-tests for significant difference. A total of 42 participants were involved in the test, in which they were divided into test and control groups. Results showed that members of the test group did not have to wait to access the test data, and accuracy rates were relatively equivalent between the groups.

X. Zhang et al. [63] proposed an automated knowledge-sharing forensic platform by applying the ontology-based approach. The proposed method involved five layers: collection, extraction, analysis, visualization, and abstraction. A timeline-based visualization panel was used to display the investigated metadata. The platform would allow forensic investigators to create schemas based on the results of their forensic investigation. However, the authors did not further discuss the applied visualization model, which might be due to the study focusing on the knowledge-sharing approach.

Chow and Ab Rahman [64] demonstrated a mobile forensic visualization tool that visualized metadata from the Android data partitions of different models. Examples of data visualization were frequent message texts, top contacts, call duration, and location maps. Tool usability was evaluated through application functionality testing and user testing. In a similar study, [51] presented a tool to visualize WhatsApp and WeChat metadata. Examples of data visualization were chat history, timeline of chats, frequent contacts, and location maps. It was observed that the visualization models and evaluation methods used in this study were similar to those of [64]. Both studies highlighted the use of forensic visualization to enhance forensic analysis tasks for mobile forensics.

Shidek et al. [65] demonstrated timeline graph visualization to display data from WhatsApp chat conversations. The utility of the tool was evaluated through questions derived from forensic analysis goals: what cyber incident occurred, who was involved in the incident, and where, when, and how the incident occurred. Also applying timeline graph visualization, [66] presented a visualization-based approach to support malware investigations on the Internet of Things environment. This study applied the data mining method to preprocess DLL files and assign weights to represent malicious and benign files. It was evaluated using a questionnaire survey for academicians and industry practitioners. The respondents were asked about visualization evaluation, user experience, and time performance. A *t*-test was used to examine the significant difference for performance evaluation.

Fig. 2 and Fig. 3 summarize the results from the literature survey of forensic analysis of IM apps. Further discussion of the findings is in the next section.

	Chang and Chang [2]	Agrawal and Tapaswi [51]	Al-Rawashdeh et al. [52]	Barradas et al. [53]	Nizam et al. [54]	Ong and Ab Rahman [55]
Forensic analysis tasks	M	M	M	A	A	A
Evaluation metrics	Origin of data, data completeness	Point of estimate value, margin of error	Count of recovered messages, average no. of recovered messages	Analysis time	Functionality testing, user acceptance testing	Functionality testing, user acceptance testing
Evaluation method	QL	QN	QN	QN	QL & QN	QL & QN
Event reconstruction/ Analysis method	N	N	N	Y /temporal	N	Y /temporal

M = manual analysis; A= automated analysis; QL = qualitative; QN = quantitative; Y = Yes; N = No

Fig. 2. Summary of studies in forensic analysis of IM apps (first part)

	Anglano et al. [44]	Riadi et al. [75]	H. Zhang et al. [47]	Rathi et al. [48]	Kim et al. [50]	Güneş Eriş and Akbal [76]
Forensic analysis tasks	M	M	M	M	M	M
Evaluation metrics	Completeness, repeatability, generality	Artifact extraction percentage, tool success rate	Origin of data, data completeness	Origin of data, data completeness	Password recovery time	Origin of data
Evaluation method	QL	QN	QL	QL	QN	QL
Event reconstruction/ Analysis method	Y / temporal and relational	N	N	N	N	Y / temporal and relational

M = manual analysis; A= automated analysis; QL = qualitative; QN = quantitative; Y = Yes; N = No

Fig. 3. Summary of studies in forensic analysis of IM apps (second part)

## VI. RESULTS AND DISCUSSION

In this section, we present the results of, and insight on, forensic analysis of Instant Messaging (IM) apps and forensic visualization for forensic analysis.

### A. Forensic Analysis of Instant Messaging Apps

The findings are summarized into methods of forensic analysis tasks, evaluation metrics, evaluation methodologies, and event reconstruction methods.

Fig. 2 and Fig. 3 present most of the studies manually analyzed forensic artifacts of IM apps, while three studies demonstrated the use of automated tools to perform forensic analysis tasks. A closer inspection of the table shows that the evaluation metrics for manual forensic analysis were different from those for automated forensic analysis. Origin of data and completeness were the most applied metrics to evaluate manual forensic analysis. For automated forensic analysis, tool performance and tool functionality were the two applied metrics.

Our examination of evaluation methodologies shows that the qualitative methodology was more predominant to be applied in manual forensic analysis studies. This is likely related to the most applied evaluation metrics. For example, the evaluation metric of origin of data is more meaningful in a qualitative way rather than quantitative. Mixed methodologies, on the other hand, are observed as the major selection of research designs evaluating automated forensic analysis tools.

It is identified that very few studies used manual analysis method and generate event reconstruction. This is an unexpected finding, since event reconstruction is the outcome of forensic analysis. A likely explanation is that extensive manual tasks might demotivate the research works from including event reconstruction in their research questions. This was echoed by Kang et al. [67], who highlighted that event reconstruction in manual analysis is limited to personal knowledge, prone to human errors, and time-consuming. In contrast, two out of the three studies that used automated forensic analysis included event reconstruction. This further supports the benefit of automating forensic analysis tasks, which is expediting event reconstruction.

As evidenced by the number of publications from 2017 until 2021, it can thus be suggested that IM forensic analysis has received significant interest from the forensics community, and the research on automated IM forensic analysis is growing. Therefore, we argue that automating IM forensic analysis to facilitate forensic analysis tasks is a promising research area. This is supported by a previous study by Anglano et al. [36], which indicated that the automated forensic analysis of mobile apps has recently received interest and that their proposed tool was able to achieve greater artifact coverage than did previous studies.

### B. Research Trends of Forensic Visualization for Forensic Analysis

The research trends were summarized based on InfoVis data types, visualization techniques, case studies, and the number of publications (see Table II). The InfoVis data types

were adopted from [53], while the list of visualization techniques was adopted from a survey conducted in [68].

TABLE II. TRENDS OF FORENSIC VISUALIZATION APPLICATIONS TO SUPPORT FORENSIC ANALYSIS

InfoVis Data Types	Forensic Visualization Techniques	Case Studies	References
Relational	Bar Chart	Network Traffic, Windows OS, Intelligent Transport System	[62], [63], [69]
	Pie Chart	Network Traffic, Electronic Mail System	[57], [60]
	Histogram	Windows OS, Electronic Mail System	[59], [60]
	Treemap / Graph	Mobile Apps (SMS), Network Traffic, Windows Diagnostic Log	[5], [57], [70]
Multi-dimensional and high-dimensional	Colorization table, RGB Binary	Container and Virtual Machine, Malware detection	[62], [71]
Text	Word cloud	Mobile Apps (text-based communication apps), Android Operating System, Instant Messaging Apps	[5], [64], [51]
Geo-spatial	Geographic maps	Mobile Apps, Android Operating System, Instant Messaging Apps	[5], [64], [51]
Sequential and temporal	Time-based/timeline	Windows OS, Cloud Computing, Internet of Things (IoT), Electronic Mail System, WhatsApp Artifact, Android Operating System, Instant Messaging Apps, Container and Virtual Machine	[64], [72], [51], [59], [60], [65], [62]

From Table II, we can see that forensic visualization was applied in various case studies, including recent computing trends, such as the Internet of Things. This finding accords with an earlier observation in [52], which showed various applications of InfoVis. Similarly, it indicates the generality of forensic visualization for incorporation into investigations of various digital infrastructures.

It is observed that each study incorporated more than one forensic visualization technique. Furthermore, most of the related studies demonstrated the selection of techniques based on data types, for example in [64], [57], [61], and [66]. This further indicates that different data types may require different visualization techniques. Therefore, we argue that the feature of the selection of visualization techniques must be incorporated in forensic visualization tools to ensure effective evidence interpretation.

Charts and graphs are the major visualization techniques applied, and relational is the most studied data type. This might be related to the usage flexibility of charts and graphs to visualize the relational data type. It can be argued that various types of charts and graphs are significant in providing evidence interaction, as well as interpretation, to users. For instance, [61] demonstrated that various parameters of drone flight data and sensor data can be visualized using various types of charts.

This is also consistent with the list of visualization techniques from a previous survey conducted in [68].

Time-based visualization is the second-most applied visualization technique. This is an unsurprising finding since timeline analysis is a major aspect of forensic investigations. The importance of timeline and its connection with digital forensics investigations were also echoed in the studies by [25] and [23]. Timeline analysis is not limited to examining the time of an incident, but it is also applicable to many other purposes, for instance, event correlation and time zone determination.

Word cloud and geographic maps were also applied as case studies involving mobile forensics data. This is, therefore, not surprising because most smartphone apps comprise textual and geo-spatial data. For example, the three case studies of [5], [64], and [51] used the word cloud to visualize textual data from communication and IM apps. This indicates that the metadata from smartphones involved enormous text-based data. Therefore, applying a word cloud can help analysts quickly understand the patterns of text-based data in relation to forensic investigations. Furthermore, incorporating geographic maps into the tool would expedite investigators in assessing geospatial data without manually examining the data using external map applications.

Another important finding is the evaluation methods of the proposed approaches. Table III summarizes the evaluation methods observed in the surveyed studies. The most applied method is performing technical simulations to evaluate tool usability, followed by conducting user testing to evaluate the tool, and using a focus group to solve forensic challenges. This shows that evaluation methods for visual analysis tools are relatively not mutually exclusive. Therefore, there is a need for benchmarking evaluation methods in this research area.

TABLE III. EVALUATION METHODS OF FORENSIC VISUALIZATION

Evaluation method	References
Technical simulations (e.g., case studies) to evaluate tool usability	[56], [59], [60], [61], [63]
User testing to evaluate tool functionality and performance	[64], [51], [62], [66]
Focus group to solve forensic challenges	[57], [58], [65]

In relation to the automated forensic tools metrics proposed by Ayers [32], speed (absolute and relative), reliability, accuracy, and completeness are the evaluation metrics used in existing studies. This is in line with our finding from automated IM forensic analysis tools, which show that performance and functionality are the most applied metrics. It should be noted that these studies might use different terms to describe the metrics. The result suggests that the existing tools were validated based on widely recognized metrics of computer forensics.

Despite the benefits of forensic visualization, most of the previous studies provided limited discussion on the selection criteria of forensic visualization techniques in relation to forensic analysis. This reflects the finding in [68] that most existing applications do not meet all forensic selection criteria. Forensic selection criteria include the following [68]: meaning, where the interpretation of evidence remains unaltered by the

visualization technique used; errors, which is the ability to identify and account for errors to prevent evidence from being questioned; transparency, which is the ability to examine and verify all the data; and timeline, which shows users the events that occurred within a specific timeframe. These criteria are important to ensure the analysis results are valid and admissible in court of law. This observation suggests a knowledge gap(s) that needs to be addressed to highlight the significance of forensic visualization.

## VII. CONCLUSION AND FUTURE WORKS

In this study, a literature survey was conducted to examine state-of-the-art IM forensic analysis and forensic visualization techniques. It appears from our literature survey that both forensic analysis and InfoVis are relatively mature research areas. However, there is a growing interest in forensic visualization and automated IM forensic analysis. This brings various research opportunities to fill up the knowledge gaps. For instance, there is a need to benchmark evaluation methods and metrics in both areas. Furthermore, there is a lack of discussion on forensic selection criteria in existing forensic visualization works.

Therefore, our next research work will be conducted on the automated forensic analysis of IM applications by integrating the forensic visualization approach. Statistical analysis and machine learning algorithms would be incorporated to utilize the forensic selection criteria in forensic visualization. It is expected that the outcomes of this study will significantly aid digital forensics practitioners in analyzing and interpreting evidence data and aid judicial authorities in understanding the presentation of evidence.

## ACKNOWLEDGMENT

The first author is currently a PhD student at Universiti Tun Hussein Onn Malaysia. This research was supported by the Ministry of Higher Education (MOHE) through Fundamental Research Grant Scheme (FRGS/1/2020/ICT07/UTHM/03/1). The authors would like to thank the anonymous reviewers for their constructive and generous feedback.

## REFERENCES

- [1] "Instant Messaging Security." <https://www.nortonlifelockpartner.com/security-center/instant-messaging-security.html> (accessed Jun. 15, 2022).
- [2] Securelist, "Kaspersky spam and phishing report for 2021 | Securelist," Kaspersky, 2021. <https://securelist.com/spam-and-phishing-in-2021/105713/> (accessed Jun. 15, 2022).
- [3] "2022 Cyber Security Statistics: The Ultimate List Of Stats, Data & Trends | PurpleSec." <https://purplesec.us/resources/cyber-security-statistics/> (accessed Jun. 15, 2022).
- [4] "WhatsApp hijack scam continues to spread - BBC News." <https://www.bbc.com/news/technology-57357301> (accessed Jun. 15, 2022).
- [5] C. F. R. Tassone, B. Martini, and K. K. R. Choo, "Visualizing Digital Forensic Datasets: A Proof of Concept," *J Forensic Sci*, vol. 62, no. 5, pp. 1197–1204, 2017, doi: 10.1111/1556-4029.13431.
- [6] H. Henseler and S. van Loenhout, "Educating judges, prosecutors and lawyers in the use of digital forensic experts," *DFRWS 2018 EU - Proceedings of the 5th Annual DFRWS Europe*, vol. 24, pp. S76–S82, 2016, doi: 10.1016/j.diin.2018.01.010.
- [7] N. D. W. Cahyani, B. Martini, and K. K. R. Choo, "Using multimedia presentations to enhance the judiciary's technical understanding of

- digital forensic concepts: An Indonesian case study," Proceedings of the Annual Hawaii International Conference on System Sciences, vol. 2016-March, pp. 5617–5626, 2016, doi: 10.1109/HICSS.2016.695.
- [8] N. D. W. Cahyani, B. Martini, and K. K. R. Choo, "Using multimedia presentations to improve digital forensic understanding: A pilot study," in ACIS 2015 Proceedings - 26th Australasian Conference on Information Systems, 2015, pp. 1–9.
- [9] W. Xiong and R. Lagerström, "Threat modeling – A systematic literature review," *Comput Secur*, vol. 84, pp. 53–69, 2019, doi: 10.1016/j.cose.2019.03.010.
- [10] N. H. Ab Rahman and K. K. R. Choo, "A survey of information security incident handling in the cloud," *Comput Secur*, vol. 49, pp. 45–69, 2015, doi: 10.1016/j.cose.2014.11.006.
- [11] H. D. Karen Kent, Suzanne Chevalier, Tim Grance, "Guide to integrating forensic techniques into incident response (NIST Special Publication 800-86)," 2006.
- [12] J. Benoit, "Best Practice Document: Forensic Analysis and Incident Handling," 2016.
- [13] E. Casey, *Digital Evidence and Computer Crime*, Second edi. San Diego, CA.: Elsevier Academic Press, 2011.
- [14] G. Horsman et al., "A forensic examination of web browser privacy-modes," *Forensic Science International: Reports*, vol. 1, p. 100036, 2019, doi: 10.1016/j.fsir.2019.100036.
- [15] B. Almaslukh, "Forensic analysis using text clustering in the age of large volume data: A review," *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 6, pp. 71–76, 2019, doi: 10.14569/ijacsa.2019.0100610.
- [16] P. Joseph and J. Norman, "Identifying Forensic Interesting Files in Digital Forensic Corpora by Applying Topic Modelling," in *Advances in Distributed Computing and Machine Learning*, Singapore: Springer, 2021, pp. 411–421.
- [17] M. Hina, M. Ali, A. R. Javed, F. Ghabban, L. A. Khan, and Z. Jalil, "SeFACED: Semantic-Based Forensic Analysis and Classification of E-Mail Data Using Deep Learning," *IEEE Access*, vol. 9, pp. 98398–98411, 2021, doi: 10.1109/ACCESS.2021.3095730.
- [18] "Computer Forensics: Forensic Techniques, Part 2 [Updated 2019] - InfosecResources," INFOSEC, 2019. <https://resources.infosecinstitute.com/topic/computer-forensics-forensic-techniques-part-2/> (accessed Nov. 21, 2021).
- [19] H. M. Elgohary, S. M. Darwish, and S. M. Elkaffas, "Improving Uncertainty in Chain of Custody for Image Forensics Investigation Applications," *IEEE Access*, vol. 10, no. 1, pp. 14669–14679, 2022, doi: 10.1109/ACCESS.2022.3147809.
- [20] I. A. Alnajjar and M. Mahmuddin, "The Enhanced Forensic Examination and Analysis for Mobile Cloud Platform by Applying Data Mining Methods," *Webology*, vol. 18, no. August 2021, pp. 47–74, 2021, doi: 10.14704/WEB/V18SI01/WEB18006.
- [21] M. Khanafseh, M. Qatawneh, and W. Almobaideen, "A survey of various frameworks and solutions in all branches of digital forensics with a focus on cloud forensics," *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 8, pp. 610–629, 2019, doi: 10.14569/ijacsa.2019.0100880.
- [22] N. A. Adderley, "Graph-based Temporal Analysis in Digital Forensics," AIR FORCE INSTITUTE OF TECHNOLOGY, 2019.
- [23] N. Adderley and G. Peterson, "Interactive temporal digital forensic event analysis," *IFIP Adv Inf Commun Technol*, vol. 589 IFIP, pp. 39–55, 2020, doi: 10.1007/978-3-030-56223-6\_3.
- [24] S. Bhandari and V. Jusas, "An ontology based on the timeline of Log2timeline and psort using abstraction approach in digital forensics," *Symmetry (Basel)*, vol. 12, no. 4, pp. 1–24, 2020, doi: 10.3390/SYM12040642.
- [25] H. Henseler and J. Hyde, "Technology assisted analysis of timeline and connections in digital forensic investigations," in *CEUR Workshop Proceedings*, 2019, vol. 2484, pp. 32–37.
- [26] H. Arshad, A. Jantan, G. K. Hoon, and I. O. Abiodun, "Formal knowledge model for online social network forensics," *Comput Secur*, vol. 89, p. 101675, 2020, doi: 10.1016/j.cose.2019.101675.
- [27] B. Carrier, *File system forensic analysis*. Addison-Wesley Professional, 2005.
- [28] A. Singh, H. S. Venter, and A. R. Ikuesan, "Windows registry harnesser for incident response and digital forensic analysis," *Australian Journal of Forensic Sciences*, vol. 52, no. 3, pp. 337–353, 2018, doi: 10.1080/00450618.2018.1551421.
- [29] B. Martini, D. Quang, and K.-K. R. Choo, "Conceptual Evidence Collection and Analysis Methodology for Android Devices," in *Cloud Security Ecosystem*, R. Ko and K. K. R. Choo, Eds. Waltham, MA: Syngress, an Imprint of Elsevier, 2015, pp. 383–400.
- [30] X. Zhang, O. Upton, N. L. Beebe, and K. K. R. Choo, "IoT Botnet Forensics: A Comprehensive Digital Forensic Case Study on Mirai Botnet Servers," *Forensic Science International: Digital Investigation*, vol. 32, p. 300926, 2020, doi: 10.1016/j.fsidi.2020.300926.
- [31] T. Wu, F. Breitingner, and S. O'Shaughnessy, "Digital forensic tools: Recent advances and enhancing the status quo," *Forensic Science International: Digital Investigation*, vol. 34, p. 300999, 2020, doi: 10.1016/j.fsidi.2020.300999.
- [32] D. Ayers, "A second generation computer forensic analysis system," *Digit Investig*, vol. 6, pp. S34–S42, 2009, doi: 10.1016/j.diin.2009.06.013.
- [33] X. Lin, T. Chen, T. Zhu, K. Yang, and F. Wei, "Automated forensic analysis of mobile applications on Android devices," *Digit Investig*, vol. 26, pp. S59–S66, 2018, doi: 10.1016/j.diin.2018.04.012.
- [34] A. Kumar, K. S. Kuppusamy, and G. Aghila, "FAMOUS: Forensic Analysis of Mobile devices Using Scoring of application permissions," *Future Generation Computer Systems*, vol. 83, pp. 158–172, 2018, doi: 10.1016/j.future.2018.02.001.
- [35] K. P. Subedi, D. R. Budhathoki, and D. Dasgupta, "Forensic analysis of ransomware families using static and dynamic analysis," *Proceedings - 2018 IEEE Symposium on Security and Privacy Workshops, SPW 2018*, pp. 180–185, 2018, doi: 10.1109/SPW.2018.00033.
- [36] C. Anglano, M. Canonico, and M. Guazzone, "The Android Forensics Automator (AnForA): A tool for the Automated Forensic Analysis of Android Applications," *Comput Secur*, vol. 88, pp. 1–15, 2020, doi: 10.1016/j.cose.2019.101650.
- [37] D. Brezinski and T. Killalea, "Guidelines for Evidence Collection and Archiving," Internet Engineering Task Force, 2002. <https://datatracker.ietf.org/doc/html/rfc3227> (accessed Jun. 01, 2022).
- [38] F. Alief, Y. Suryanto, L. Rosselina, and T. Hermawan, "Analysis of autopsy mobile forensic tools against unsent messages on whatsapp messaging application," in *International Conference on Electrical Engineering, Computer Science and Informatics (EECSI)*, 2020, pp. 26–30. doi: 10.23919/EECSI50503.2020.9251876.
- [39] C. Anglano, M. Canonico, and M. Guazzone, "Forensic analysis of Telegram Messenger on Android smartphones," *Digit Investig*, vol. 23, pp. 31–49, 2017, doi: 10.1016/j.diin.2017.09.002.
- [40] A. Akinbi and E. Ojie, "Forensic analysis of open-source XMPP/Jabber multi-client instant messaging apps on Android smartphones," *SN Appl Sci*, vol. 3, no. 4, pp. 1–14, 2021, doi: 10.1007/s42452-021-04431-9.
- [41] A. F. Imam Riadi, "Forensic Analysis of Android Based Instant Messaging Application," in *12th International Conference on Telecommunication Systems, Services, and Applications (TSSA)*, 2018, pp. 1–6.
- [42] H. Zhang, L. Chen, and Q. Liu, "Digital Forensic Analysis of Instant Messaging Applications on Android Smartphones," *2018 International Conference on Computing, Networking and Communications, ICNC 2018*, pp. 647–651, 2018, doi: 10.1109/ICCNC.2018.8390330.
- [43] K. Rathi, U. Karabiyik, T. Aderibigbe, and H. Chi, "Forensic analysis of encrypted instant messaging applications on Android," *6th International Symposium on Digital Forensic and Security, ISDFS 2018 - Proceeding*, vol. 2018-Janua, pp. 1–6, 2018, doi: 10.1109/ISDFS.2018.8355344.
- [44] R. Chanajitt, W. Viriyasitavat, and K. K. R. Choo, "Forensic analysis and security assessment of Android m-banking apps," *Australian Journal of Forensic Sciences*, vol. 50, no. 1, pp. 3–19, 2018, doi: 10.1080/00450618.2016.1182589.
- [45] G. Kim, S. Kim, M. Park, Y. Park, I. Lee, and J. Kim, "Forensic analysis of instant messaging apps: Decrypting Wickr and private text messaging

- data,” *Forensic Science International: Digital Investigation*, vol. 37, p. 301138, 2021, doi: 10.1016/j.fsidi.2021.301138.
- [46] M. S. Chang and C. Y. Chang, “Forensic analysis of LINE messenger on android,” *Journal of Computers (Taiwan)*, vol. 29, no. 1, pp. 11–20, 2018, doi: 10.3966/199115992018012901002.
- [47] V. Agrawal and S. Tapaswi, “Forensic analysis of Google Allo messenger on Android platform,” *Information and Computer Security*, vol. 27, no. 1, pp. 62–80, 2019, doi: 10.1108/ICS-03-2017-0011.
- [48] A. M. Al-Rawashdeh, Z. A. Al-Sharif, M. I. Al-Saleh, and A. S. Shatnawi, “A Post-Mortem Forensic Approach for the Kik Messenger on Android,” 2020 11th International Conference on Information and Communication Systems, ICICS 2020, pp. 79–84, 2020, doi: 10.1109/ICICS49469.2020.239559.
- [49] D. Barradas, T. Brito, D. Duarte, N. Santos, and L. Rodrigues, “Forensic analysis of communication records of messaging applications from physical memory,” *Comput Secur*, vol. 86, pp. 484–497, 2019, doi: 10.1016/j.cose.2018.08.013.
- [50] S. H. S. Nizam, N. H. Ab Rahman, and N. D. W. Cahyani, “Keyword Indexing And Searching Tool (KIST): A Tool to Assist the Forensics Analysis of WhatsApp Chat,” *International Journal on Information and Communication Technology (IJoICT)*, vol. 6, no. 1, p. 23, 2020, doi: 10.21108/ijoict.2020.61.481.
- [51] W. S. Ong and N. H. Ab Rahman, “A Forensic Analysis Visualization Tool for Mobile Instant Messaging Apps,” *International Journal on Information and Communication Technology (IJoICT)*, vol. 6, no. 2, pp. 78–87, 2020, doi: 10.21108/IJoICT.2020.00.530.
- [52] D. Rees and R. S. Laramee, “A survey of information visualization books,” *Computer Graphics Forum*, vol. 38, no. 1, pp. 610–646, 2019.
- [53] M. Behrisch et al., “Quality Metrics for Information Visualization,” *Computer Graphics Forum*, vol. 37, no. 3, pp. 625–662, 2018, doi: 10.1111/cgf.13446.
- [54] M. Sorapure, “Text, Image, Data, Interaction: Understanding Information Visualization,” *Comput Compos*, vol. 54, p. 102519, 2019, doi: 10.1016/j.compcom.2019.102519.
- [55] I. Krak, O. Barmak, and E. Manziuk, “Using visual analytics to develop human and machine-centric models: A review of approaches and proposed information technology,” *Comput Intell*, vol. 1, no. 9, pp. 75–98, 2020, doi: 10.1111/coin.12289.
- [56] F. Böhm, L. Englbrecht, and G. Pernul, “Designing a decision-support visualization for live digital forensic investigations,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 12122, Springer, 2020, pp. 223–240. doi: 10.1007/978-3-030-49669-2\_13.
- [57] I. Kotenko, M. Kolomeets, A. Chechulin, and Y. Chevalier, “A visual analytics approach for the cyber forensics based on different views of the network traffic,” *J Wirel Mob Netw Ubiquitous Comput Dependable Appl*, vol. 9, no. 2, pp. 57–73, 2018, doi: 10.22667/JOWUA.2018.06.30.057.
- [58] D. Clark and B. Turnbull, “Interactive 3D visualization of network traffic in time for forensic analysis,” *VISIGRAPP 2020 - Proceedings of the 15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, vol. 3, pp. 177–184, 2020, doi: 10.5220/0008950601770184.
- [59] S. K. Weng and J. Y. Tu, “A visualization jump lists tool for digital forensics of windows,” *KSII Transactions on Internet and Information Systems*, vol. 14, no. 1, pp. 221–239, 2020, doi: 10.3837/tiis.2020.01.013.
- [60] R. Lu and L. Li, “Research on forensic model of online social network,” 2019 IEEE 4th International Conference on Cloud Computing and Big Data Analytics, ICCCBDA 2019, pp. 116–119, 2019, doi: 10.1109/ICCCBDA.2019.8725746.
- [61] A. Renduchintala, F. Jahan, R. Khanna, and A. Y. Javaid, “A comprehensive micro unmanned aerial vehicle (UAV/Drone) forensic framework,” *Digit Investig*, vol. 30, no. 2019, pp. 52–72, 2019, doi: 10.1016/j.diin.2019.07.002.
- [62] J. Shropshire and R. Benton, “Container and VM visualization for rapid incident response,” in *Proceedings of the Annual Hawaii International Conference on System Sciences*, 2020, pp. 6397–6406.
- [63] X. Zhang, K. K. R. Choo, and N. L. Beebe, “How Do i Share My IoT Forensic Experience with the Broader Community? An Automated Knowledge Sharing IoT Forensic Platform,” *IEEE Internet Things J*, vol. 6, no. 4, pp. 6850–6861, 2019, doi: 10.1109/IJOT.2019.2912118.
- [64] C. X. Quan and N. H. Ab Rahman, “A Mobile Forensic Visualization Tool for Android Data Partition,” in *Applied Information Technology And Computer Science*, 2021, vol. 2, no. 2, pp. 37–52.
- [65] H. Shidek, N. D. W. Cahyani, and A. A. Wardana, “WhatsApp Chat Visualizer: A Visualization of WhatsApp Messenger’s Artifact Using the Timeline Method,” *International Journal on Information and Communication Technology (IJoICT)*, vol. 6, no. 1, pp. 1–9, 2020, doi: 10.21108/ijoict.2020.61.489.
- [66] I. Ahmad, M. A. Shah, H. A. Khattak, Z. Ameer, M. Khan, and K. Han, “FIViz: Forensics investigation through visualization for malware in internet of things,” *Sustainability (Switzerland)*, vol. 12, no. 18, pp. 1–23, 2020, doi: 10.3390/SU12187262.
- [67] J. Kang, S. Lee, and H. Lee, “A digital forensic framework for automated user activity reconstruction,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 7863, Berlin, Heidelberg: Springer, 2013, pp. 263–277. doi: 10.1007/978-3-642-38033-4\_19.
- [68] C. Tassone, B. Martini, and K. K. R. Choo, “Forensic Visualization: Survey and Future Research Directions,” in *Contemporary Digital Forensic Investigations of Cloud and Mobile Applications*, Syngress, 2017, pp. 163–184. doi: 10.1016/B978-0-12-805303-4.00011-3.
- [69] D. Gürdür and L. Söpjani, “Visual Analytics to Support the Service Design for Sustainable Mobility,” 2018 IEEE Conference on Technologies for Sustainability, SusTech 2018, pp. 4–9, 2019, doi: 10.1109/SusTech.2018.8671353.
- [70] S. Park and S. Lee, “DiagAnalyzer: User behavior analysis and visualization using Windows Diagnostics logs,” *Forensic Science International: Digital Investigation*, vol. 43, p. 301450, Sep. 2022, doi: 10.1016/j.fsidi.2022.301450.
- [71] O. J. Falana, A. S. Sodiya, S. A. Onashoga, and B. S. Badmus, “Mal-Detect: An intelligent visualization approach for malware detection,” *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 5, pp. 1968–1983, May 2022, doi: 10.1016/j.jksuci.2022.02.026.
- [72] S. C. Sathe and N. M. Dongre, “Data acquisition techniques in mobile forensics,” in *Proceedings of the 2nd International Conference on Inventive Systems and Control, ICISC 2018*, 2018, pp. 280–286. doi: 10.1109/ICISC.2018.8399079.

# Driving Maneuvers Recognition and Classification Using A Hybrid Pattern Matching and Machine Learning

Munaf Salim Najim Al-Din

Department of Electrical and Computer Engineering, University of Nizwa, Birkat Al-Mouz, Nizwa, P.C. 616, Oman

**Abstract**—Since most of the road and traffic accidents are related to human errors or distraction, the study of irregular driving behaviors is considered one of the most important research topics in this field. To prevent road accidents and assess driving competencies, there is an urgent need to evaluate driving behavior through the design of a driving maneuvers assessment system. In this study, the recognition and classification of highway driving maneuvers using smartphones' built-in sensors are presented. The paper examines the performance of three classical machine learning techniques and a novel hybrid system. The proposed hybrid system combines the pattern matching Dynamic Time Warping (DTW) technique for recognizing driving maneuvers and the machine learning techniques for classification. Results obtained from both approaches show that the performance of the hybrid system is superior to that obtained by using classical machine learning techniques. This enhancement in the performance of the hybrid system is due to the elimination of the overlapping in the target classes due to the separation, the recognition and the classification processes.

**Keywords**—Driving behavior classification; driving maneuvers; pattern matching; machine learning

## I. INTRODUCTION

According to previous studies in the field of traffic safety and road accidents, abnormal or irregular driving behaviors have been considered to be one of the main factors that greatly contribute to road accidents [1]. With the increase of vehicles all over the world, abnormal driving patterns detection and monitoring will most defiantly contribute to the reduction of road accidents. In addition to the above benefits, studies of driving patterns and behaviors have been instrumental in the development of advanced driver assistance systems (ADAS) and autonomous vehicles (AVs) [2, 3]. Driving behaviors can be assessed from two different perspectives namely; drivers' actions or the vehicle's dynamic state. In the first approach the driver is considered as the focal element where a set of parameters that affect the driver's vigilance and attention are continuously observed to predict his/her competence to achieve the driving course in a robust and safe manner [4]. Drivers' state monitoring systems may contain different modules, such as facial recognition systems, physiological signals monitoring and drivers' interaction and control. For example, drivers' interaction and control, combined with facial recognition, have been shown to be effective in detecting driver fatigue, drowsiness, and distraction [4].

In the second approach, the dynamic state of the vehicle, such as longitudinal and lateral accelerations, braking, is monitored to detect and classify abnormal driving patterns or maneuvers. In general, signals captured through the vehicle's built-in sensors captured through the CAN-BUS [5, 6], or external sensors such as accelerometers and gyroscopes, and GPS [7, 8], or a combination of in-vehicle and external sensors [9], can be used for the aforesaid purposes. In the past ten years, smartphones have emerged as an efficient and very reliable tool in this field, since they have powerful computational capabilities, richness and variety of built-in sensing devices and ability to have multiple ways of communication with external devices connected to the OBD-II port. Furthermore, smartphones especially with the emergence of 5G technology have been enabled to play cooperative coordinator between vehicles through vehicle-to-everything networks. With all the above listed features provided by smartphones, attention has been immensely focused on the utilization of smartphones in monitoring and analyzing driving behaviors.

The analysis of driving behavior is dependent on the maneuvers to be analyzed as well as the collected data or estimated parameters used to describe them. Various methods were proposed in the literature to perform this task. The simplest approach considers the driving process as a rule-based or fuzzy classification problem. A set of thresholds are defined or extracted, based on experience or trial-and-error, to assess the driving parameters and then classify driving maneuvers [7, 10-18]. In general, these methods are not reliably accurate because the thresholds, fuzzy sets and rules, as well as the classification results, are all based on presuppositions. The second approach is based on pattern matching and recognition techniques, such as Dynamic Time Warping (DTW) [19]. This approach is based on measuring the level of similarity between captured signals and standard patterns. The disadvantage of using classical DTW is the heavy computational burden especially when dealing with multivariate time series.

Recent research has demonstrated that machine learning techniques are capable of identifying and classifying irregular driving patterns using models and rules that evaluate driving maneuvers and then driving behavior. Machine learning approaches are generally classified into supervised learning approaches and unsupervised learning approaches. Various supervised learning approaches, such as K-Nearest Neighbor (KNN), Naive Bayes, Decision Trees, and Random Forest (RF)

[20], linear regression [21], Support Vector Machines (SVM) [22, 23], and Neural Networks [24] require the extraction of features, such as statistical values, time domain parameters, and frequency domain parameters, for training. On the other hand, unsupervised learning approaches, such as K-means clustering [25] and Principal Component Analysis algorithms [21] can infer and generate rules and threshold-based discriminators for clustering purposes. During the past decade, different methodologies and techniques have been proposed and implemented successfully in the field of driving behavior classification [4, 8].

In this paper, three classical machine learning techniques namely Random Forest (RF), Support Vector Machine (SVM) and K-Nearest Neighbor (KNN) were used to recognize and classify six highway driving maneuvers. The data required for training and testing the machine learning models were collected through smartphones' accelerometers and gyroscope sensors. Furthermore, a novel hybrid approach based on the integration of pattern matching DTW and machine learning approaches has also been proposed and investigated in this study. The basic idea of this hybrid approach is to separate the recognition process from the classification process. The DTW developed in this study is used to provide signal similarity measures for the input signals, while the three above-mentioned machine learning techniques were utilized for classification.

The rest of the paper is organized as follows: Section II introduces briefly three machine learning techniques, the RF, the SVM and the KNN. Section III provides a brief description of the structure and workflow of the system. In Section IV, the maneuver detection unit is described with emphasis on the implementation of an adaptive sliding window. In section V, the structure and implementation of the driving maneuvers identification unit is presented. Evaluation of the performance of the two approaches is presented and compared in Section VI. The conclusions are drawn in Section VII.

## II. CLASSICAL MACHINE LEARNING

A wide range of techniques have been developed to recognize and classify driving maneuvers in the literature. In recent years, driving maneuver classification using machine learning techniques has received increasing attention for the evaluation of driving patterns and drivers' profiling [4]. Three machine learning techniques have been used in this study, namely RF, SVM, and KNN, for recognizing and classifying driving maneuvers. These three techniques will be discussed briefly:

### A. Random Forest Technique

A random forest classifier is an ensemble classifier that is made up of a set of decision trees trained on different sub-sets of the training data and then their predictions are aggregated to improve prediction accuracy and control over-fitting. An RF classifier usually uses bootstrap aggregation and boosting, in which random samples of the training dataset are selected with replacement and trained independently. The use of bagging and feature randomness to generate a set of decision tree classifiers typically results in high variance and low correlation. As a solution to this problem, these decision trees are usually

connected in parallel and by using majority voting the variance is minimized and thus the prediction is improved. The implementation of RF classifier is summarized as follows: [26]:

- 1) Select M random samples from the labeled training set using the bootstrapping technique.
- 2) Construct a RF with N parallel decision trees.
- 3) Form N samples to train the N parallel decision tree models as follows:
  - a) For each feature x in a given feature set  $N_i$  calculate the Information gain from the entropy of the classes and the entropy of the feature x.
  - b) Find the node with the maximum information gain and split it into sub-nodes.
  - c) Iterate through a and b to form the tree until reaching the lowest amount of samples needed to split.
- 4) Repeat steps (1) and (2) to get N tree classifiers.
- 5) For testing data, find the prediction of each decision tree, and allocate the new data to the category that wins the majority votes using the following formula:

$$P^*(x) = \max_y \sum_{t=1}^N I(N_t(x) = P(x)) \quad (1)$$

In the formula,  $P^*(x)$  is the classification result of random forest,  $N_t(x)$  is the classification result of each classification tree,  $P(x)$  is a classification target, and  $I(\cdot)$  is an indicator function which returns 1 if the condition in the argument is true, 0 otherwise.

### B. Support Vector Machine

The main function of the SVM algorithm is to find the finest hyperplane in an N-dimensional space that separates the data and clusters them based on classes by using a kernel function. The SVM is in fact a binary classifier but can be extended to handle multi-class classification by training a series of binary SVMs or by solving a single optimization problem. A high classification rate can be achieved if the optimal selected hyperplane has the largest functional margin. This margin is represented by the distance of the hyperplane to the nearest training data points of any class. For the learning process of the SVM algorithm, constrained nonlinear optimization is used to obtain an optimal hyperplane. In general, a SVM classifier uses a nonlinear mapping function that maps the data into a high-dimensional feature space to distinctly classify the data points as follows:

$$P(x) = \sum_i \lambda_i K_i \langle x \cdot x_i \rangle + b \quad (2)$$

Where,  $\lambda_i$  is support vector,  $x_i$  is data sample,  $i = 1, 2, \dots, C$ ; C number of classes and  $K_i \langle x \cdot x_i \rangle$  are a set of kernel functions defined by:

$$K_i \langle x \cdot x_i \rangle = \begin{cases} e^{-z_1 h(x)^2} & \text{if } x \in X^1 \\ \vdots & \vdots \\ e^{-z_c h(x)^2} & \text{if } x \in X^C \end{cases} \quad (3)$$

In the above equation  $h(x)$  is a binary decision function expressed as:

$$h(x) = \sum_j \lambda_j y_j \langle x \cdot x_j \rangle + b \quad (4)$$

While  $x_i$  is the  $i^{\text{th}}$  sample of the training dataset, which includes the  $N$  number of samples with  $C$  categories and the value of the parameter  $z_j$  can be computed from the chi-square test [27]. The final classification decision is made according to a rule of the form:

$$P^*(x) = \arg \max_c (w_c K_i \langle x \cdot x_i \rangle + b) \quad (5)$$

The weighting factor appearing in Eqn. 8 is defined as:

$$w_i = \frac{N/n_i}{\sum_{i=1}^C N/n_i} \quad (6)$$

Where,  $N$  and  $C$  denote the training sample size and category size, respectively.  $n_i$  indicates the sample size of every category with  $i = 1, 2, \dots, C$ .

The implementation of the SVM classifier is summarized as follows:

- 1) Select  $M$  random samples from the labeled training set using the 5-fold technique and initialize the kernel matrix  $K_i$ .
- 2) For each sample  $x$  calculate:
  - a) Calculate the distance  $h_j(x)$ ,  $\{j = 1, 2, \dots, C\}$ ;  $C$  number of classes.
  - b) Calculate the value of the weighting factor  $w_j$  and parameter  $z_j$  for every support vector.
- 3) Find  $P(x)$  from Eqn.(2).
- 4) Find the new Kernel matrix from  $P(x)$  and from the previous Kernel matrix.
- 5) Repeat steps (2) to (4) until finding the optimal hyperplane, i.e  $h_j(x)$  with optimal functional margin.
- 6) For testing data, find the prediction from equation (4).

### C. K-Nearest-Neighbors Technique

The KNN is a supervised machine learning algorithm for classifying classes based on their feature similarity to other classes. In the KNN the classification of a certain testing sample depends on its distance with respect to other samples in the training dataset. The distance between two samples is employed to measure their similarity [28]. The distance is calculated using different measures such as the Chebyshev distance, the Euclidean distance, and more generally the Minkowski distance. In this paper the Minkowski distance between two feature vectors is used. Where the Minkowski distance is a distance measured between two points in  $N$ -dimensional feature space by the following formula:

$$d(x_i - x_j) = \left( \sum_{i=1}^C |x_i - x_j|^p \right)^{\frac{1}{p}} \quad (7)$$

Where  $x_i$  and  $x_j$  are two features vectors and  $p$  is an integer value.

The implementation of the KNN classifier is summarized as follows:

- 1) Select  $M$  random samples from the labeled training set using the 5-fold technique.
- 2) Set the value of the nearest data points  $K$  which can be any integer preferably to be odd integer.
- 3) For every point in the testing data do the following:

- a) Compute the distance between the test data and each sample in the training data as in Eqn(7).
- b) Sort the distances obtained in (a) in an ascending order.
- c) Select the first  $K$  rows from the sorted distances array.
- d) Assign a class to the test point depending on most frequent class of these rows.

### III. SYSTEM STRUCTURE

Fig. 1, shows the general workflow of the proposed system. The system consists of four main interrelated units namely, data collection unit, data processing unit, maneuver recognition unit, and finally, maneuver classification unit. In this section, the functions of the first two units are briefly introduced. A detailed description of the operation of these units can be found in [29].

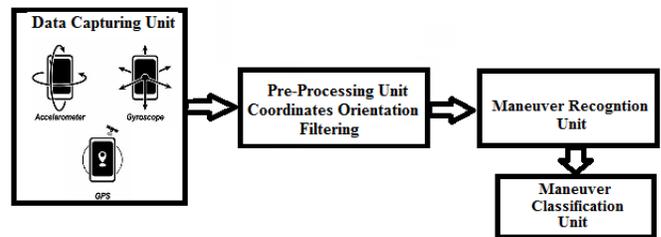


Fig. 1. The proposed system workflow.

Using calibrated Android smartphones with built-in accelerometers and gyroscopes, raw vehicle data was collected at a rate of 50 samples/second. The calibration method for the smartphones' IMUs sensors are adopted from [30]. As well as the data captured by the IMUs, the smartphones' GPS data was used for referencing the location of the vehicle.

The pre-processing unit is intended to achieve two main functions namely, signal filtering, and transformation of sensors data to the vehicle's coordinate system. The first problem is typically attributed to the fact that the IMUs in Smartphones are based on MEMS technology, thus they suffer from white Gaussian noise. Furthermore, the sensors are very sensitive hence they capture in addition to the variation of the dynamic parameters of the vehicle's vibration [30]. Fig. 2 shows instantaneous captured data for a sample maneuver. A locally weighted running line smoother (LOSS) filtering technique is used for removing this noise and smoothing the recorded signals. The use of this type of filter was investigated and its performance was compared with two other filters; the one-dimensional Kalman filter and the simple moving average filter. The LOSS filter was found to be the best effective filtering approach when compared with others, and Fig. 3 shows a sample of a smoothed signal [29-30].

A coordinate reorientation module is integrated with the pre-processing to correct the collected sensors' data by aligning the smartphone's coordinate system with the vehicle coordinate system. By presuming that the vehicle is driven on a horizontal road, during the initial calibration, the vehicle roll and pitch angles relative to a tangent frame both can be considered to be zero. Furthermore, if the vehicle does not experience any

acceleration, the smartphone's roll and pitch angles can be estimated from accelerometer measurements of the gravity vector. This can be done using a set of geometrical rotations using Euler angles. The determination of Euler angles is fully explained in [31].

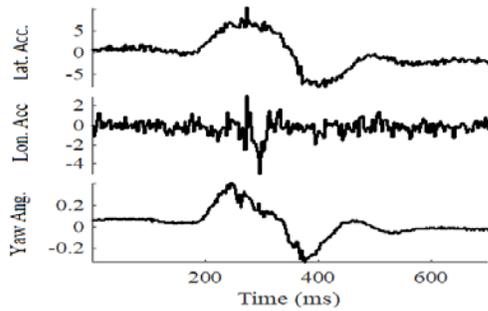


Fig. 2. Raw signals captured by the smartphone's IMUs.

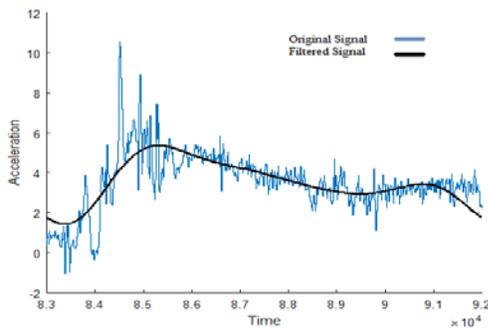


Fig. 3. Raw and filtered accelerometer signal.

#### IV. MANEUVERS DETECTION

Table I, presents the list of maneuvers that can be detected by the proposed system. These maneuvers have been detected by an adaptive sliding window with a short-term energy endpoint detection algorithm.

TABLE I. MANEUVERS CLASSES

1-	Acceleration straight road segment	2-	Braking straight road segment
3-	Left lane change straight road segment	4-	Right lane change straight road segment
5-	Merging into highway	6-	Exit from highway

Maneuvers are detected in three iterative stages. In the first stage a window of 100ms width is used to compute the short-term energy of the signal. Based on the fact that for an infinite sequence of a discrete signal the energy is defined by:

$$\hat{y}[n] = y[n]W[m - n], m - M + 1 \leq n \leq M \quad (8)$$

Where  $W$  is a window function given by:

$$W[m - n] = \begin{cases} 1 & 0 \leq m \leq M - 1 \\ 0 & \text{Otherwise} \end{cases} \quad (9)$$

The energy contained in this short interval then can be computed by:

$$E_w = \sum_{n=m-M+1}^m (y[m]W[n - m])^2 \quad (10)$$

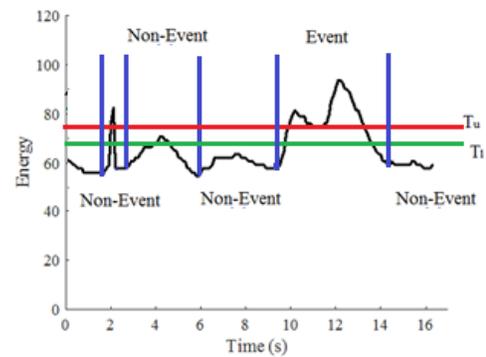


Fig. 4. Maneuver detection using short-term energy.

Once the short-energy is computed it will be compared with a set of pre-defined thresholds, as shown in Fig. 4. If this energy is less than a specific threshold  $T_l$  for the whole 100msec window, then this frame will be ignored and will be considered a non-event. Otherwise, if the energy is greater than  $T_l$ , then the starting time of the event detected is recorded and the short-term energy is computed for a sliding window as in Eq. (10). The width of the window will increase by 20msec and the short-term energy will be calculated over the whole interval of the extended window. For each step in this stage the following conditions will be checked:

- If the computed short-term energy remains less than the upper threshold  $T_u$  for 1 second or drops below  $T_l$  in a short time, then this segment will be considered a false event and the system will start with a new 100msec window as in the first stage.
- When the short-term energy for the extendable sliding window is computed to be higher than the upper threshold  $T_u$  for more than one second, then the system will consider this signal as a result of an event. If the system records the starting time of the event, it will continue to compute the short-term energy for the extendable sliding window and compare it with  $T_u$ . If the short-term energy drops below  $T_l$  for more than 100ms, the system will record the ending time of the event.

#### V. MANEUVERS IDENTIFICATION

Generally, supervised machine learning techniques such as decision trees, support vector machines, neural networks, and many others are used to identify and classify types of driving maneuvers in a single process. All these techniques require a set of features to represent the input signals such as time, frequency or statistical features, for training and testing. In this study, sixteen time and statistical features listed in Table II were used to train and test the recognition and classification performance of the first approach. It should be noted that classical machine learning techniques, when trained with time and statistical features, cannot provide a clear description of how patterns of signals behave. In this regard, it is difficult to draw any conclusions from the parameters of the systems. Additionally, errors resulting from recognition and classification will accumulate and affect the performance of these techniques.

TABLE II. SELECTED STATISTICAL FEATURES

1	Mean	9	Peak to peak value
2	Median	10	Peak to RMS value
3	Maximum value	11	Root-sum-of-squares
4	Minimum value	12	Skewness
5	Standard of deviation	13	Kurtosis
6	Mode	14	Range of values
7	Variance	15	Interquartile range values
8	RMS value	16	The mean absolute deviation

Due to the fact that time-varying signals are required to recognize the types of driving maneuvers, it has been shown that pattern recognition or matching techniques, such as the DTW technique, are superior in this regard. DTW identifies the types of maneuvers by comparing input patterns against standard templates and calculating the similarity level between them. As a result of using the DTW method, incoming signals can be compared with a predefined standard template regardless of any differences in their amplitudes or durations. Therefore, it would be likely to have a set of standard templates to measure the similarity of maneuvers for different drivers [32-35]. It should be noted that the main disadvantage of the DTW approach is its extreme computational requirements, since it computes the similarity level between all the possible patterns in the input signals. In the case that multi-signal identification is required, this problem will become more complex. Furthermore, a considerable amount of work is required to select and compute the reference templates because it is very difficult to collect all possible templates that would cover all driving styles and behaviors of drivers [19].

Fig. 5 shows the basic structure of the DWT unit. The DTW technique utilizes discrete dynamic programming to determine the similarity between two signals, regardless of any difference in time, frequency, or deformation related effects to dynamic spatiotemporal differences. In a previous study [29], the recognition unit was implemented using  $(n \times m)$  DTW units, where  $(n)$  represents the number of signals and  $(m)$  represents the number of standard patterns for each signal. As a consequence, for every detected event, i.e. driving maneuver, a  $(n \times m)$  matrix containing warping cost is derived by comparing all the signals with all the stranded templates. This study has reduced the amount of computation required by the classical DTW technique by reducing the number of signals used to recognize driving maneuvers, as well as by utilizing energy activation units. In this study, the implementation of the DWT technique is based on two facts, which have been demonstrated in previous studies. The first is that there are only three signals, longitudinal acceleration, lateral acceleration, and yaw angle. The second fact is that the signals vary according to certain patterns, so their energy depends on these variations, see Table III.

When using the DWT technique to identify the type of any signal, a set of standard signals, or templates, are required to compare the unknown input signal with them and measure the similarity level. The selection of these reference signals for each specific class is not a straight forward task since the set of the collected signals, for each maneuver class, have different

time durations and amplitudes. There are three different approaches in electing a suitable reference signal from a set of measurements namely; the longest common sequence approach, the medoid sequence approach, and the average sequence approach. In this paper, for a specific DWT unit, the signal that has the minimum average of distances with all the signals in that set is extracted and elected to be the reference signal or template. The details of this novel methodology are given in [35].

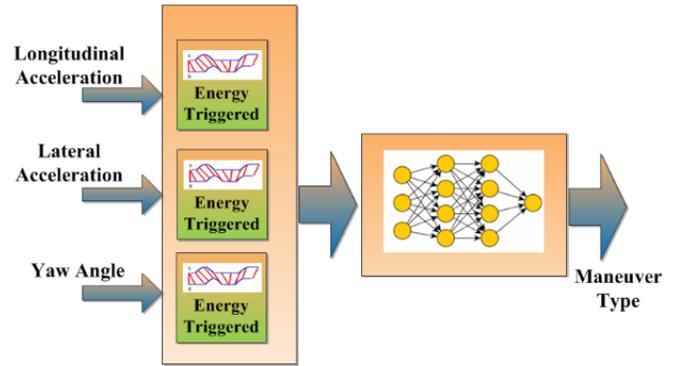


Fig. 5. Maneuvers recognition unit.

TABLE III. TYPICAL SIGNALS' PATTERNS FOR DIFFERENT MANEUVER

Maneuver Type	Longitudinal Acceleration	Lateral Acceleration	Yaw Angle
Acceleration			
Break			
Left-Lane Change			
Right-Lane Change			
Exit			
Merge			

## VI. EXPERIMENTS AND RESULTS

In this work two different approaches were used to recognize and classify highway driving maneuvers. The first approach utilizes the classical machine techniques described in Section II, while in the second approach the DTW and the aforementioned techniques were integrated to create a hybrid system. With this system, the DTW method will be used for the

recognition process while classical machine learning techniques will be applied for the classification process.

### A. Experimental Data

Before exploring the analysis and results, it is worth mentioning that the development of the system progressed through two levels, the development level and the naturalistic driving testing level.

At the development level ten drivers with different types of vehicles and experience were volunteered to drive through a 16km highway road segment that has different configurations and conditions as shown in Fig. 6. Each driver was asked to execute the driving maneuvers listed in Table I with different categories; i.e. Light, Normal and Hard. All of the vehicles were equipped with smartphones that were programmed to collect sensor data at a rate of 50Hz and four cameras that recorded the surrounding vehicles. Every class of driving maneuver was performed by each driver at least five times, so the total number of driving maneuvers gathered in this phase was 900 samples. This part of the dataset was then presented to experts to obtain their judgment and to build the knowledge base that is required for labeling the maneuvers. This initial dataset was used to train and test the two suggested systems.



Fig. 6. The route used to collect initial dataset.

The 5-fold cross-validation technique was used from which 60% of the initial dataset were utilized to generate and extract the data and the features that are required in the computation of the DTW reference templates, define the lower and upper limits that define the range of values of each cluster, i.e. class and statistical features vectors for training the systems.

### B. Models Evaluation

To assess the validation of the two approaches the remaining 40% of the initial dataset has been used to validate their performance.

The first assessment of the system was to test its capability to detect driving maneuvers, i.e. recording the starting and ending time. Fig. 7 shows a portion of a short trip conducted to cover some of the basic maneuvers. As shown in Fig. 7, these are the raw data that were captured directly from the calibrated smartphone's sensors. Fig. 8 illustrates the pre-processed signals, i.e. after smoothing the signals of Fig. 7. As shown in the figure, the red rectangles represent the output of the maneuver detection unit. As it can be seen, the unit effectively detects the beginning and the end of any variation in the input signals. According to the testing of maneuver detection unit with manually registered maneuvers, the detection rate was more than 96%.

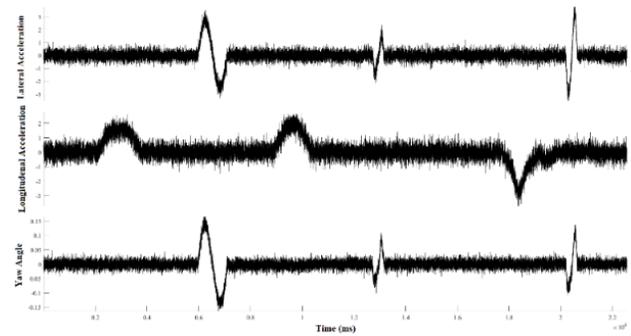


Fig. 7. Raw data captured from the smartphone's sensors.

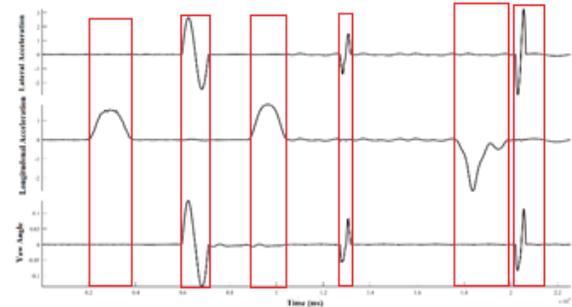


Fig. 8. Signals after filtration and maneuver detection.

Three evaluation metrics namely Precision (PR), Recall (RC), and F1-score (F1) have been used for evaluating the performance of each system in addition to the confusion matrix. Precision is generally defined as the probability that a certain class of maneuvers is correctly classified in either recognition or classification results. In contrast, recall is the probability that all maneuvers in a particular maneuver class are correctly identified. Finally, the F1-score is determined based on both precision and recall, as shown in Eq. (11), where a high F1-score indicates the system's overall performance quality.

$$F1 = \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2TP}{2TP + FP + FN} \quad (11)$$

Where TP is true positive, FP is false positive and FN is false negative. All these three values can be found using the confusion matrix.

In the first approach, namely the three classical machine learning models, all the statistical features listed in Table III were obtained for each segmented maneuver. It should be noted that the models are performing both the recognition and classification processes. The confusion matrix for the RF model is shown in Fig. 9, and Table IV presents a comparison for each maneuver of the three models in terms of PR, RC and F1.

As it can be seen from Table IV, the performance of the RF model is the highest, where the average precision of the model is 0.84, the recall is 0.833 and the F1 score is 0.835. For the SVM the parameters are (PR = 0.783, RC = 0.772, F1 = 0.775) and for the KNN they are (PR = 0.749, RC = 0.736, F1 = 0.74). It is not an easy task to dig for the actual factors behind the low performance of the models when compared with the RF.

However, both SVM and KNN are not efficient algorithms when they deal with large data sets, and they do not function well when the target classes are overlapped. The RF model is able to handle large datasets because it is based on the bagging algorithm which generates as many trees as possible based on the testing data and generates an output combining the tree outputs. Therefore, the RF techniques can be considered as an ensemble learning approach, hence it would reduce the overfitting problem in decision trees, reduces the variance and improves the accuracy.

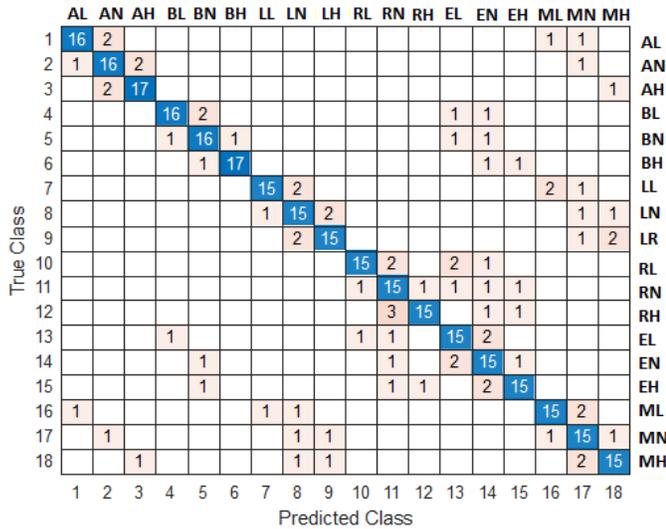


Fig. 9. Confusion matrix for the classical RF implementation.

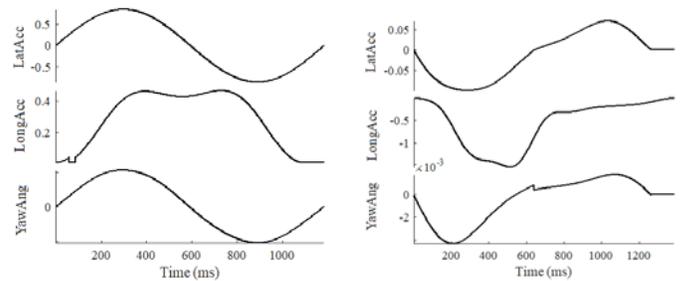
TABLE IV. COMPARISON OF THE MODELS FIRST APPROACH

Class	RF			SVM			KNN		
	PR	RC	FI	PR	RC	FI	PR	RC	FI
AL	0.85	0.85	0.85	0.89	0.80	0.84	0.83	0.75	0.79
AN	0.81	0.85	0.83	0.76	0.80	0.78	0.71	0.75	0.73
AH	0.94	0.85	0.89	0.85	0.85	0.85	0.84	0.80	0.82
BL	0.89	0.85	0.87	0.89	0.80	0.84	0.88	0.75	0.81
BN	0.77	0.85	0.81	0.76	0.80	0.78	0.71	0.75	0.73
BH	0.95	0.90	0.92	0.94	0.85	0.89	0.88	0.75	0.81
LL	0.89	0.85	0.87	0.88	0.75	0.81	0.83	0.75	0.79
LN	0.84	0.80	0.82	0.68	0.75	0.71	0.67	0.70	0.68
LH	0.89	0.85	0.87	0.79	0.75	0.77	0.78	0.70	0.74
RL	0.89	0.80	0.84	0.88	0.75	0.81	0.83	0.75	0.79
RN	0.76	0.80	0.78	0.65	0.75	0.70	0.68	0.75	0.71
RH	0.89	0.80	0.84	0.88	0.75	0.81	0.82	0.70	0.76
EL	0.81	0.85	0.83	0.68	0.75	0.71	0.70	0.70	0.70
EN	0.67	0.80	0.73	0.60	0.75	0.67	0.54	0.75	0.63
EH	0.89	0.80	0.84	0.79	0.75	0.77	0.70	0.70	0.70
ML	0.85	0.85	0.85	0.79	0.75	0.77	0.79	0.75	0.77
MN	0.70	0.80	0.74	0.63	0.75	0.68	0.56	0.70	0.62
MH	0.81	0.85	0.83	0.75	0.75	0.75	0.71	0.75	0.73

It should be mentioned here that a thorough analysis has been conducted in this study to identify the overlap in the target classes. It was found that there are two groups of maneuvers which could have a high similarity rate between their classes. The first group contains the Acceleration, Left-Lane change and the Merging maneuvers and the second group contains the other three maneuver classes. Fig. 10(a) illustrates a signal that was manually recorded as a left-lane change, while the system recognized it as a merging maneuver. On the other hand, Fig. 10(b) shows a break maneuver but has been recognized by the system as an exit maneuver. From the point of view of the author, this noise in the dataset needs careful analysis, hence it will be left to a future investigation.

In the second approach the same 5-fold cross-validation method was used to extract the DTW reference templates and again to train and validate the same models but for a specific maneuver type. As mentioned previously in this approach the DTW unit is acting as a recognition unit while the three classical machine learning models are acting as classifiers.

The performance of the DTW was first tested and it was found that the structure of the unit needs some modification to overcome the problem of overlapping classes. A simple two-hidden layers neural network was integrated into the unit, where the three measured distances obtained from each DTW are fed as an input to this neural network. Fig. 11 shows the confusion matrix for the predicted maneuvers. All the performance measures, precision, recall and F1-score were calculated for the recognition unit and they are equal to 0.95, which indicates an excellent validity of the recognition unit.



(a): LLC recognized merge. (b): Break recognized exit.

Fig. 10. Examples for misrecognized classes.

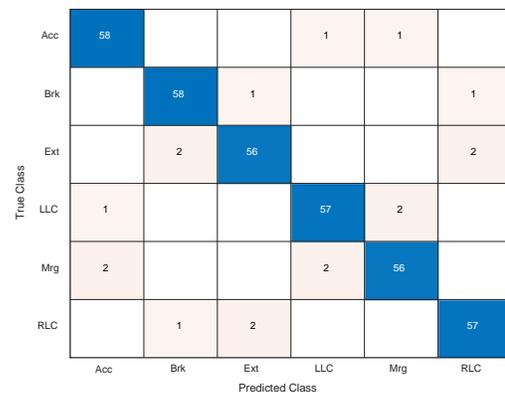


Fig. 11. Confusion matrix for the recognition unit.

Table V presents a comparison between the three models that perform a classification process for each maneuver separately. Fig. 12 shows samples of the confusion matrix for different cases. Again the performance of the RF model is the highest when compared with the others and still the KNN model has the lowest performance. The average precision of the RF model is 0.908, the recall is 0.905 and the F1 score is 0.91. These newly obtained results indicate an enhancement of 9% is achieved when using the second approach. Similar improvements were also noticed in the other models, where for the SVM the performance indicators are PR = 0.875, RC = 0.871, F1 = 0.87 and an average enhancement of 12.25%, while the performance indicators for the KNN model are PR = 0.838, RC=0.835, F1 = 0.84 and an average enhancement of 13.5%.

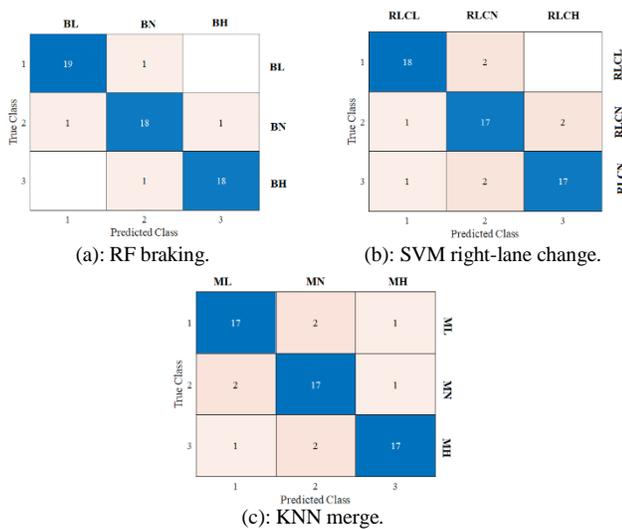


Fig. 12. Confusion matrices samples for hybrid system initial dataset.

TABLE V. COMPARISON OF THE MODELS SECOND APPROACH

Class	RF			SVM			KNN		
	PR	RC	F1	Class	PR	RC	F1	Class	PR
AL	0.95	0.90	0.92	0.89	0.89	0.89	0.85	0.89	0.87
AN	0.86	0.90	0.88	0.81	0.85	0.83	0.80	0.80	0.80
AH	0.95	0.95	0.95	0.95	0.90	0.92	0.89	0.85	0.87
BL	0.95	0.95	0.95	0.90	0.90	0.90	0.86	0.90	0.88
BN	0.90	0.90	0.90	0.81	0.85	0.83	0.80	0.80	0.80
BH	0.95	0.95	0.95	0.95	0.90	0.92	0.89	0.85	0.87
LL	0.95	0.90	0.92	0.90	0.90	0.90	0.85	0.85	0.85
LN	0.82	0.90	0.86	0.81	0.85	0.83	0.80	0.80	0.80
LH	0.95	0.90	0.92	0.89	0.85	0.87	0.85	0.85	0.85
RL	0.95	0.95	0.95	0.90	0.90	0.90	0.85	0.85	0.85
RN	0.86	0.90	0.88	0.81	0.85	0.83	0.76	0.80	0.78
RH	0.95	0.90	0.92	0.89	0.85	0.87	0.89	0.85	0.87
EL	0.90	0.90	0.90	0.89	0.85	0.87	0.84	0.80	0.82
EN	0.81	0.85	0.83	0.74	0.85	0.79	0.70	0.80	0.74
EH	0.95	0.90	0.92	0.94	0.85	0.89	0.89	0.80	0.84
ML	0.95	0.90	0.92	0.90	0.90	0.90	0.85	0.85	0.85
MN	0.78	0.90	0.84	0.81	0.85	0.83	0.81	0.85	0.83
MH	0.94	0.85	0.89	0.95	0.90	0.92	0.89	0.85	0.87

### C. Naturalistic Driving Testing

In the second stage of this study, a comprehensive dataset was collected by installing only the data collection app onto the smartphones of 25 drivers who drove frequently to and from various locations and the University of Nizwa, as shown in Fig. 13. Data collected in this phase is real naturalistic driving data based on different routes that are very dynamic and include many different types of roads.



Fig. 13. The routes used for collecting the naturalistic dataset.

After performing the necessary preprocessing for the data of each driver, the captured signals were analyzed by an off-line Matlab code to detect and extract driving actions by using the adaptive sliding window described in Section V. Table VI provides a list of the number of driving events obtained in this phase.

TABLE VI. COMPARISON OF THE MODELS SECOND APPROACH

Type of Maneuver	#	Type of Maneuver	#
Acceleration	680	Right-Lane Change	495
Braking	595	Exit	256
Left-Lane Change	508	Merge	255

As a result of the high number of maneuvers obtained from smartphone sensors, a separate module was developed to label the maneuvers in addition to the suggested system. The module uses a semi-supervised labeling system based on the DTW technique. The module is similar to the DTW recognition unit with the exception that it was specifically designed to identify maneuver classes. The only difference between the two systems is that there are nine DTW units and each one is devoted to a single class of a certain maneuver. A detailed explanation of the implementation of this technique can be found in [36]. The distance between two time series signals is given by:

$$Distance(A, B) = \frac{DTW(A, B)}{ED(A, B) + \delta} \quad (12)$$

Where A is a standard reference signal, or template used by the DTW and Euclidian distance calculation, B is the signal that needs to be classified, DTW(A, B) is the distance measured by the classical constrained DTW algorithm, ED(A, B) is the classical Euclidian distance and d is an extremely small positive quantity used to avoid divide-by-zero error.

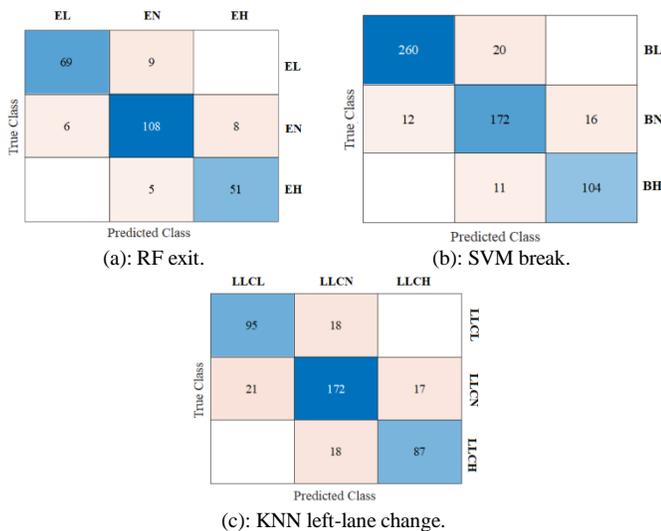


Fig. 14. Confusion matrices samples for hybrid system naturalistic dataset.

Fig. 14 shows the confusion matrix for different cases. As it has been expected the RF model has the highest performance with respect to the SVM and KNN, while the KNN model is still showing the lowest performance. The average precision, recall and F1-score are all approximately 0.9, those for the SVM are 0.87 and finally those for the KNN are 0.834. As it can be seen, the results are almost the same for both datasets and this gives a positive indication that the suggested approach is stable and reliable.

## VII. CONCLUSIONS

Two different approaches are presented in this paper for the recognition and classification of highway driving maneuvers using smartphone sensors. Raw data captured through smartphone's IMUs sensors are first pre-processed by transforming sensors' data from the smartphone's coordinates system to the actual vehicle coordinates system, then these data were smoothed by using the LOSS filter and finally, the longitudinal and lateral acceleration and the yaw angle are deduced from these data. Three parameters were found to be sufficient to recognize and classify driving maneuvers.

The first approach investigated in this paper utilizes three different classical machine learning techniques, namely RF, SVM and KNN techniques. Results obtained from this approach showed that RF had the highest performance when compared to SVM and KNN. This superiority of the RF model can be attributed to the fact that the RF model can handle large datasets efficiently. It's based on the bagging algorithm and uses the Ensemble Learning technique. Nevertheless, it was found that the classical implementation of machine learning techniques suffers from a serious problem in dealing with noisy data, i.e. overlapping in the target classes. It was found that there are two groups of maneuvers which could have a high similarity rate between their classes. The first group contains the Acceleration, Left-Lane change and the Merging maneuvers and the second group contains the other three maneuver classes.

In this paper, a hybrid technique is used to overcome the overlapping between the classes. The recognition unit of this

approach utilizes a novel DTW unit that demonstrates an excellent recognition rate with F1-Score of 0.95. The maneuver classifications are then obtained by machine learning techniques. When compared to the classical approach, the performance of the novel approach was significantly improved.

A large dataset was collected from naturalistic driving for 25 drivers on different highways. About 2800 maneuvers were obtained from this dataset. With such a high number of maneuvers a semi-supervised labeling system based on the DTW technique was used. The module is similar to the DTW recognition unit but was trained solely for labeling maneuver classes. The second approach was tested on the second dataset. Results obtained show a high rate of recognition and classification, nearly the same as that obtained with the first dataset.

## REFERENCES

- [1] A. Haghi, D. Ketabi, M. Ghanbari and H. Rajabi, "Assessment of Human Errors in Driving Accidents; Analysis of the Causes Based on Aberrant Behaviors", *Life Science Journal*, vol 11, No. 9, pp. 414-420, 2014.
- [2] Muhammad Qasim Khan and Sukhan Lee, "A Comprehensive Survey of Driving Monitoring and Assistance Systems", *Sensors*, vol. 19, issue 11, pp. 2574-2606, 2019.
- [3] Clara Marina Martinez, Mira Heucke, Fei:Yue Wang, Bo Gao, Dongpu Cao, "Driving Style Recognition for Intelligent Vehicle Control and Advanced Driver Assistance: A Survey", *IEEE Transactions on Intelligent Transportation Systems*, vol. 19, issue 3, pp. 666-676, 2018.
- [4] Dengfeng Zhao, Yudong Zhong, Zhijun Fu, Junjian Hou, Mingyuan Zhao, "A Review for the Driving Behavior Recognition Methods Based on Vehicle Multisensor Information", *Journal of Advanced Transportation*, vol. 2022, pp. 1-16, 2022.
- [5] R. Kridalukmana, H. Y. Lu and M. Naderpour, "An object oriented Bayesian network approach for unsafe driving maneuvers prevention system", in the 2017 12th International Conference on Intelligent Systems and Knowledge Engineering (ISKE), pp. 1-6, 2017.
- [6] Iván Silva and JoséEugenio Naranjo, "A Systematic Methodology to Evaluate Prediction Models for Driving Style Classification", *Sensors*, vol. 20, issue 6, pp. 1692-1713, 2020.
- [7] H. Malik, G. S. Larue, A. Rakotonirainy and F. Maire, "Fuzzy Logic to Evaluate Driving Maneuvers: An Integrated Approach to Improve Training," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 4, pp. 1728-1735, 2015.
- [8] T. K. Chan, C. S. Chin, H. Chen and X. Zhong, "A Comprehensive Review of Driver Behavior Analysis Utilizing Smartphones," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 10, 4444-4475, 2020.
- [9] N. AbuAli, "Advanced vehicular sensing of road artifacts and driver behavior," the 2015 IEEE Symposium on Computers and Communication (ISCC), pp. 45-49, 2015.
- [10] L. M. Bergasa, D. Almería, J. Almazán, J. J. Yebes and R. Arroyo, "DriveSafe: An app for alerting inattentive drivers and scoring driving behaviors" the 2014 IEEE Intelligent Vehicles Symposium Proceedings, pp. 240-245, 2014.
- [11] M. Fazeen, B. Gozick, R. Dantu, M. Bhukhiya and M. C. González, "Safe Driving Using Mobile Phones," *IEEE Transactions on Intelligent Transportation Systems*, vol. 13, no. 3, pp. 1462-1468, pp. 1462:1468, 2012.
- [12] Yang Zheng, A. Sathyanarayana and J. H. L. Hansen, "Threshold based decision-tree for automatic driving maneuver recognition using CAN-Bus signal," the 2014 17th International IEEE Conference on Intelligent Transportation Systems (ITSC), pp. 834-2839, 2014.
- [13] T. Pholprasit, W. Choochaiwattana and C. Saiprasert, "A comparison of driving behaviour prediction algorithm using multi-sensory data on a smartphone," the 2015 IEEE/ACIS 16th International Conference on

- Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD), pp. 1-6, 2015.
- [14] Hamid Reza Eftekhari and Mehdi Ghatee, "Hybrid of discrete wavelet transform and adaptive neuro fuzzy inference system for overall driving behavior recognition", *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 58, pp. 782-796, 2018.
- [15] Hamid Reza Eftekhari, Mehdi Ghatee.: "A similarity:based neuro:fuzzy modelling for driving behavior recognition applying fusion of smartphone sensors", *Journal of Intelligent Transportation Systems*, vol. 23, no. 1, pp. 72-83, 2019.
- [16] C. Arroyo, L. M. Bergasa and E. Romera, "Adaptive fuzzy classifier to detect driving events from the inertial sensors of a smartphone," the 2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC), pp. 1896-1901, 2016.
- [17] A. Aljaafreh, N. Alshabat and M. S. Najim Al-Din, "Driving style recognition using fuzzy logic," the 2012 IEEE International Conference on Vehicular Electronics and Safety (ICVES 2012), pp. 460-463, 2012.
- [18] Munaf Najim Al-Din, Ahmad Aljaafreh, Nashat Albdour and Maen Saleh., "Driving Styles Recognition Using Decomposed Fuzzy Logic System", *International Journal of Electrical, Electronics and Computer Systems*, vol. 16, issue 1, pp. 820-824, 2013.
- [19] Munaf S. Najim Al-Din. "Real-Time Identification and Classification of Driving Maneuvers using Smartphone." *Advances in Science, Technology and Engineering Systems Journal*, vol. 5, pp. 193-205, 2020.
- [20] Wu, M.; Zhang, S.; Dong, Y., "A Novel Model-Based Driving Behavior Recognition System Using Motion Sensors", *Sensors*, vol. 16, issue 10, pp. 1746-1769, 2016.
- [21] Chen Chen, Xiaohua Zhao, "Driver's Eco-Driving Behavior Evaluation Modeling Based on Driving Events", *Journal of Advanced Transportation*, vol. 2018, pp. 1-12, 2018.
- [22] J. Yu, Z. Chen, Y. Zhu, Y. Chen, L. Kong and M. Li, "Fine-Grained Abnormal Driving Behaviors Detection and Identification with Smartphones," *IEEE Transactions on Mobile Computing*, vol. 16, no.8, pp. 2198-2212, 2017.
- [23] J. F. J´unior, E. Carvalho, B. V. Ferreira, C. de Souza, Y. Suhara, A. Pentland, and G. Pessin, "Driver behavior profiling: An investigation with different smartphone sensors and machine learning", *PLOS one*, vol. 12, no. 4, e0174959, 2017.
- [24] P. Brombacher, J. Masino, M. Frey and F. Gauterin, "Driving event detection and driving style classification using artificial neural networks," the 2017 IEEE International Conference on Industrial Technology (ICIT), pp. 997-1002, 2017.
- [25] U. Fugiglando et al., "Driving Behavior Analysis through CAN Bus Data in an Uncontrolled Environment," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 2, pp. 737-748, 2019.
- [26] Breiman, L., "Random Forests," *Machine Learning*, vol. 45, pp. 5-32, 2001.
- [27] Ketu, S., Mishra, P.K., "Scalable kernel-based SVM classification algorithm on imbalance air quality data for proficient healthcare", *Complex Intell. Syst.* vol. 7, pp. 2597-2615, 2021.
- [28] S. Ray, "A Quick Review of Machine Learning Algorithms", the 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon), pp. 35-39, 2019.
- [29] Munaf Salim Najim Al-Din; Atef Saleh Al-Mashakbeh, "Development of a highway driving events identification and classification using smartphone", *International Journal of Nanoparticles*, vol. 12, issue 1-2, pp. 152-173, 2020.
- [30] M. S. Najim Al-Din, "Calibration and Pre-Processing Techniques for a Smartphone-Based Driving Events Identification and Classification System," the 2018 IEEE Electron Devices Kolkata Conference (EDKCON), pp. 396-402, 2018.
- [31] D. A. Johnson and M. M. Trivedi, "Driving style recognition using a smartphone as a sensor platform," the 2011 14th International IEEE Conference on Intelligent Transportation Systems (ITSC), pp. 1609-1615, 2011.
- [32] Saiprasert, C., Pholprasit, T. and Thajchayapong, S., "Detection of Driving Events using Sensory Data on Smartphone", *International Journal of Intelligent Transportation Systems Research*, vol. 15, no. 1, pp. 17-28, 2015.
- [33] Gurdit Singh, Divya Bansal and Sanjeev Sofat, "A smartphone based technique to monitor driving behavior using DTW and crowdsensing", *Pervasive and Mobile Computing*, vol. 40, pp. 56-70, 2017.
- [34] Pavlo Tkachenko, Jinwei Zhou, Davide Gagliardi and Luigidel Re, "On-line Maneuver Identification in Highway Traffic Using Elastic Template Matching", *IFAC:PapersOnLine*, vol. 51, no. 15, pp. 557-562, 2018.
- [35] F. Petitjean, J. Inglada, and P. Gancarski, "A global averaging method for dynamic time warping, with applications to clustering," *Pattern Recogn.*, vol. 44, no. 3, pp. 678-693, 2011.
- [36] Yanping Chen, Bing Hu, Eamonn Keogh, and Gustavo E.A.P.A Batista., "DTW-D: time series semi-supervised learning from a single example", In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '13)*, pp. 383-391, 2013.

# An Approach to Automatic Garbage Detection Framework Designing using CNN

Akhilesh Kumar Sharma<sup>1\*</sup>, Antima Jain<sup>2</sup>, Deevesh Chaudhary<sup>3</sup>, Shamik Tiwari<sup>4</sup>, Hairulnizam Mahdin<sup>5</sup>, Zirawani Baharum<sup>6,\*</sup>, Shazlyn Milleana Shaharudin<sup>7</sup>, Ruhaila Maskat<sup>8</sup>, Mohammad Syafwan Arshad<sup>9</sup>

SIT, Manipal University Jaipur, Jaipur, Rajasthan<sup>1,3\*</sup>,

Delhi Skill and Entrepreneurship University, Rajokari Campus, Delhi, India-110038<sup>2</sup>

Virtualization Department-School of Computer Science, University of Petroleum and Energy Studies, Dehradun, Uttarakhand, India<sup>4</sup>

Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia<sup>5</sup>

Technical Foundation-Malaysian Institute of Industrial Technology, Universiti Kuala Lumpur Persiaran Sinaran Ilmu, Bandar Seri Alam, 81750 Johor Bahru<sup>6</sup>

Department of Mathematics-Faculty of Science and Mathematics, Universiti Pendidikan Sultan Idris, Tanjung Malim, Perak, Malaysia<sup>7</sup>

Department of Statistics, Columbia University, New York, N.Y., USA<sup>7</sup>

Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA Shah Alam, Selangor Malaysia<sup>8</sup>

MZR Global Sdn Bhd, 5A, Jalan Kristal K7/K, Seksyen 7, 40000 Shah Alam, Selangor, Malaysia<sup>9</sup>

**Abstract**—This paper proposes a system for automatic detection of litter and garbage dumps in CCTV feeds with the help of deep learning implementations. The designed system named Greenlock scans and identifies entities that resemble an accumulation of garbage or a garbage dump in real time and alerts the respective authorities to deal with the issue by locating the point of origin. The entity is labelled as garbage if it passes a certain similarity threshold. ResNet-50 has been used for the training purpose alongside TensorFlow for mathematical operations for the neural network. Combined with a pre-existing CCTV surveillance system, this system has the capability to hugely minimize garbage management costs via the prevention of formation of big dumps. The automatic detection also saves the manpower required in manual surveillance and contributes towards healthy neighborhoods and cleaner cities. This article is also showing the comparison between applied various algorithms such as standard TensorFlow, inception algo and faster-r CNN and Resnet-50, and it has been observed that Resnet-50 performed with better accuracy. The study performed here proved to be a stress reliever in terms of the garbage identification and dumping for any country. At the end of the article the comparison chart has been shown.

**Keywords**—Garbage detection; Resnet; TensorFlow; CNN

## I. INTRODUCTION

India generates 62 million tons of waste every year, of which less than 60% is collected i.e. around 25 million tons of waste remains in the open and forever stagnating. The country is moving towards a technological revolution, work on smart cities and sanitation have moved in to the limelight. Still, the method of collection of waste products is ancient and inefficient. Due to mismanagement of waste in urban and rural areas, garbage dumps tend to pile up around corners of vicinities creating a source of bad odor and a breeding ground for pests and diseases. Often, it is very difficult to find these accumulations of waste and eliminate them before they cause harm to a colony, neighborhood, locality or a city. The

efficiency and effectiveness of pre-existing waste control measures needs to be improved to save on monetary resources utilized in these procedures and preserve the aesthetic value of cities. An even more pressing concern is the deterioration of healthy living conditions due to these accumulations. The idea was to create a product that uses the live feed of CCTV cameras on the streets to identify the areas where garbage has begun to accumulate. This will assist the garbage collection and waste control authorities like the Nagar Nigam in cleaning the city in a less time consuming and focused fashion.

Urban garbage monitoring, currently, is done entirely with the help of human resources. In present times all major cities have CCTV cameras in place nowadays which have been put in for ensuring security, detecting and catching traffic violators and perpetrators of criminal activity. Progress has already started to be seen in the direction of automating the processing of these video feeds with computer vision instead of manual monitoring. In better developed cities in the country, automatic vehicular number plate capturing and image capturing have been devised to catch traffic offenders exceeding the speed limit without the efforts of an on-field traffic police officer. Similarly, this proposition is to viably use these camera feeds for detecting garbage dumps with the help of machine learning techniques based in computer vision. This idea is very feasible as it does not require changes in the existing infrastructure, only nominal additions to deal with the processing needs of the software to analyze the video feed, making it not only easier to implement but economically viable as well.

Large advancements have already been made in the field of using machine learning techniques using computer vision with the concepts of deep learning and neural networks (CNN) [1][2]. Frameworks based on these technologies are making automation easier as they make the need for human intervention redundant in day-to-day tasks. Neural networks are used to mimic the human mind to act in place of

individuals and focus human resources on much more important tasks. Convolutional Neural Networks are used to classify images or parts of images into differentiable categories by detecting various features of the image like edges, colors and backgrounds to classify an object present in it. Our objective is to provide a smart solution for garbage collection that analyses live video feeds and highlights garbage present in frames.

This paper is organized as follows, section one is on introduction, section two focused on literature review, section three is on the methodology which will introduce the algorithm; the experiment that will be conducted is introduced in section four and lastly section five is a conclusion.

## II. LITERATURE REVIEW

As per Huang, Kevin Murphy & Wu et al. [3], their work provided a model to select a detection architecture that obtains right speed/memory/accuracy balance for a given application and platform. Various ways were investigated to trade accuracy for speed and memory usage in modern convolutional object detection systems

As per Anitya, Kumar & Wu et al. [4], their paper introduces a Quick Locale based Convolutional System strategy (Quick R-CNN) [13][14] for question recognition. Quick R-CNN expands on past work to effectively characterize object recommendations utilizing profound convolutional systems. Quick R-CNN trains the VGG16 arrange nine times more efficiently than R-CNN, is 213 times faster at test-time, and accomplishes a higher Guide on PASCAL VOC 2012.

As per Nurminen, Jukka & Wu et al. [5] in their research work modern machine learning based approaches are used, in particular, the Yolo neural network system, to detect high-level objects, e.g. pumps or valves, in diagrams which can be scanned from paper archives or stored in vector or pixel form. In this concept, a simulator is used to automatically generate labeled training material to the system. A previously trained network is retrained to detect the components of interest.

As per Chen, Chunlin & Wu et al. [6], their research works to solve the problem of detecting objects, especially small objects, in complex scenes, the authors of this paper have proposed a novel module named as Adaptive Convolution Block (ACB), which adjusts the parameters of convolutional filters corresponding to the current feature maps and then filter these feature maps with the obtained convolutional filters to generate boosted features. Due to such adaptive convolution, the boosted features can pay more emphasis on the exercised objects, suppressing the background information caused by irrelevant surroundings and improve the detection accuracy.

As per Huang, Lili & Wu et al. [7], their research focuses on instruction-guided object detection, i.e., predicting the objects associated with the implementation of a specific instruction for intelligent robots. An amendment to the current detection paradigm is proposed by incorporating semantic instruction description effectively. To address the challenges related to picking out instruction-related objects from the detection results of a general object detector, a flexible dataset is introduced that can well adapt to the variation of the instruction set and only annotates instruction-related object

samples. An amendment to the current detection paradigm is proposed by incorporating semantic instruction description effectively.

As per Kharinov, Mikhail & Wu et al. [8], their paper focuses on the problem of automatized object detection in color images. The explication basing on the classic pixel clustering methods is discussed and advanced. The parameter for the heterogeneity of image areas is introduced. New methods for improving the quality of an image with automatically produced object names are suggested.

As per Liu, Ying & Wu et al. [9], their paper proposes a garbage detection system based on deep learning and narrowband IOT[16][17][18] concepts. The system automatically identifies garbage directly in embedded monitoring module, and manages thousands of monitoring front-ends via background server [12][25][26][27]and narrow-band Internet of Things. The improved YOLOv2 network model[28] is adopted to do garbage detection and recognition, in the front-end embedded module of the system

As per Hu, et al. [10], their research proposes a method for object detection by receiving a user input that specifies one or more first regions and one or more second regions in a template image. The other regions include furthermore objects of interest [19][20][21]. The method further constitutes each of the one or more first regions recognizing a third region in an image under detection comparable to the first region in the template image by matching the image from the point of espial with the template image [22][23][24]. The method further constitutes computing a transformation function based on the similarities from each of the one or more first regions to its comparable third region. The method further constitutes the computed transformation function to the one or more second regions to localize one or more fourth regions in the image for the object detection.

## III. METHODOLOGY

Table I shows the comparisons between CNN Algorithms which were implemented until ResNet50 was finalized as the final framework. Fig. 1 shows the general progression of the research methodology discussed in this paper.

TABLE I. CNN ALGORITHMS COMPARISON TABLE

CNN Algorithms	Confidence %
Standard Tensorflow Algo	75
Inception v1.0	85
Faster r-CNN	90
ResNet50	95

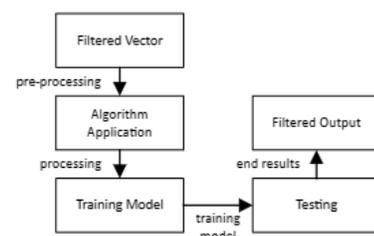


Fig. 1. Flowchart depicting the methodology

A. Algorithm Used

Learning algorithm – Learning algorithm of ResNet is main back-propagation algorithm furthermore, continuous and online learning algorithm response is applied.

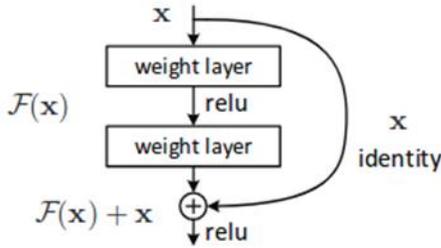


Fig. 2. Residual learning

- The identity shortcuts (x) [29] can be directly used when the input and output are of the same dimensions as shown in Fig. 2.

$$Y = F(X, \{W_i\}) + X \quad (1)$$

- When the dimensions change, A) The shortcut still performs identity mapping, with extra zero entries padded with the increased dimension. B) The projection shortcut is used to match the dimension (done by 1\*1 conv) using the following formula [29]:

$$Y = F(X, \{W_i\}) + W_s X \quad (2)$$

- The first case adds no extra parameters, the second one adds in the form of  $W\{s\}$  [29].
- ResNet50 is 3 layers deep as shown in Fig. 3.

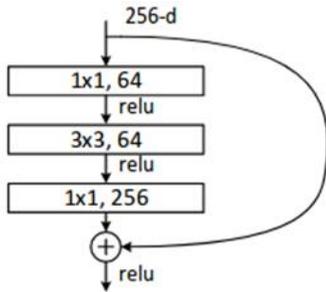


Fig. 3. ResNet50 layer architecture [29]

1) Evaluation Method

$$Recall = \frac{TP}{(TP+FN)} \quad (3)$$

$$Precision = \frac{TP}{(TP+FP)} \quad (4)$$

In this formula, TP, FP, and FN are the number of true cases, false positive cases, and false negative cases respectively.

B. Detection Model

The solution proposed in this research work boasts a minimalistic architecture thus does not have many technical requirements. Anyone can run the software to its maximum efficiency as it requires a live video source as its primary and only external component.

The research work began with intensive research into the necessary fields with the help of different research papers dealing with detection for custom objects, and several blogs detailing similar projects ideas to find the most efficient technology stack to be used. Initially work started with Tensorflow to run our basic object detection code with its library of trained objects, but the confidence threshold of the resultant matched objects was never above 50%. To circumvent this issue a private python code was used which allowed for the use of ResNet50 [29] (A convolutional neural network) model increasing the confidence percentage for more pixelated images. Running the code on these packages gave positive results. Residual learning provides the advantages like potentially deeper, thus this study uses this approach.

Next moving forward, the primary focus was on creating and training the main detection model with a custom dataset. The dataset is formed of 20000 images of various forms of garbage heaps and dumps found all over the world with plans to update the source code to implement a self-learning agent which learns from the outside environment and makes better predictions over time. A Nvidia graphics card along with tensorflow-gpu library were used for training. The trained dataset was then tested for image files. The threshold in the start was 85% but the restrains made the threshold percentage to go up to 95%. Finally, custom designed python modules derived from the raccon dataset by Datitran were used for running the webcam live feed from video source devices, using the OpenCV package of python [11]. Multiple webcams feed can work in cohesion to create a community network to prevent illicit garbage collection throughout a whole smart community. Via the output feed, the target location will automatically get pinged on the software’s hub portal and all active organizations can be informed. The whole process was done in an Anaconda environment running Python 3.5 and Tensorflow 1.12.0.

C. Technology

ResNet-50 is a convolutional neural network that is trained on more than a hundred million images from the ImageNet database. This network is 50 layers deep and can classify images into 1000 object categories. As a result, ResNet-50 has learned rich feature representations for a wide range of images and has an image input size of 224-by-224.

TensorFlow is a free and open-source software library for dataflow and differentiable programming across a range of task and a symbolic math library. It is used for machine learning applications such as neural networks. Experience in TensorFlow is a standard expectation of the machine learning industry. Several python packages are being used here, such as SciPy, NumPy, OpenCV3, Pillow, Matplotlib and Keras.

D. Modules

Identification of correct Technology/Tools – The project started with a search for the correct tools required to materialize the project’s vision. Tensorflow was the first library to be selected for the training models followed by several Python packages like OpenCV, Keras, Pillow.

Testing Models-Several models like Inceptionv3 and R-CNN were tried before settling for ResNet50 framework for

the training of the module. ResNet50, as shown in Fig. 4, when used with Tensorflow was able to provide faster training speed as well as better accuracy.

Collection of Garbage Datasets-Initially the project requirement included a large dataset consisting mainly of Indian/local street garbage and open dumps. The final dataset consists of images taken from SpotGarbage Dataset, some public datasets from Kaggle as well as a large number of personal inclusions. The dataset contains over 15000 training images and over 5000 testing images.

Training the Dataset-To train a robust detection classifier the images in the dataset were individually labelled. The image .xml data, containing information such as borders and dimensions of objects in the images, was used to create .csv files containing all the data for the train and test images in tabular structures for labelling. With the images labeled, the TFRecords were generated that served as input data to the TensorFlow training model. Xml\_to\_csv.py and generate\_tfrecord.py scripts from Dat Tran's Raccoon Detector dataset were used with slight modifications to the original work to include our custom directory structure. Then a label map was created and editing of the training configuration file was done. Then the training code was executed and it took about eight hours and two million steps for the model to be successfully trained.

Coding image/video modules and module for live feeds – The coding of the different modules was done in python programming language mainly utilizing the OpenCV package and the trained dataset in an anaconda environment running via a command line prompt.

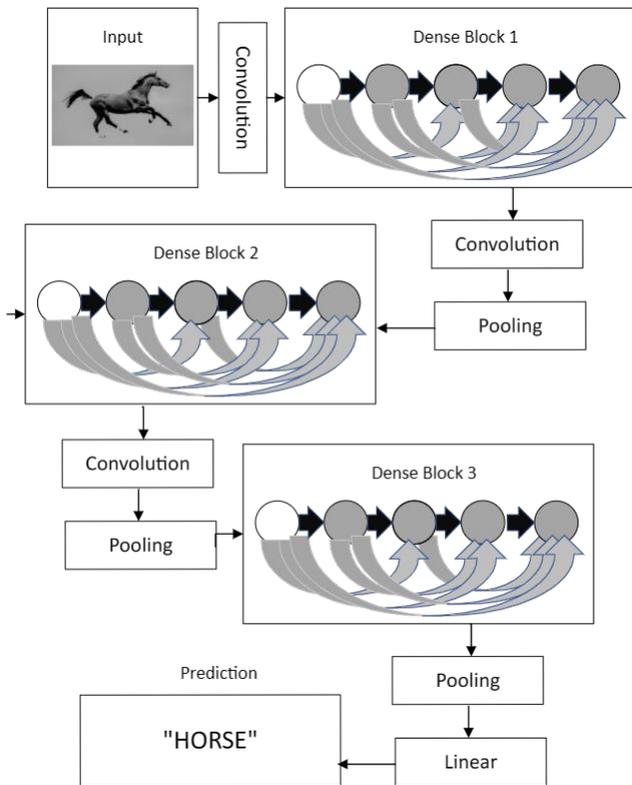


Fig. 4. A standard ResNet50 network

Testing of the Dataset- The trained module can be utilized to detect garbage in an image, video or a live webcam feed. Table II shows the summary of the test results.

TABLE II. SUMMARY OF TEST RESULTS

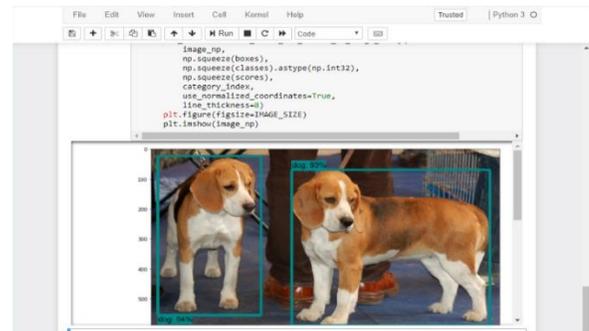
Number of test pictures	5000
Test Speed	0.71
Average Detection Accuracy	85.07
Detection Accuracy Variance	0.089

#### IV. RESULTS

Fig. 5 shows the prototype used in the experiments. The first prototype worked with object detection for all preexisting libraries of Tensorflow only on images. It can be seen in Fig. 5(a) that both the objects are being detected as dogs along with a confidence percentage. The second prototype worked with the designed custom garbage database running classifier on images, videos. It can be seen in Fig 5(b) that separate dumps of garbage are also shown as separate labels rather than just one. The third prototype worked with the designed custom garbage database running classifier on live feed. This is the finalized product, as it can be seen below in Fig. 5(b), the garbage dump present in the frame running on a live footage can be segmented very precisely and is detected with a very high certainty level. The comparative algorithmic confidence percentage is shown in Table III below.

TABLE III. COMPARATIVE ALGORITHMIC % CONFIDENCE

CNN Algorithms	Confidence %
Standard Tensorflow Algo	75
Inception v1.0	85
Faster r-CNN	90
ResNet50	95



(a)



(b)



Fig. 5. Prototype (a) 1, (b) 2 and (c) 3

In Fig. 6, it has been observed that the total loss during training and the chart shows the decrement of the loss with respect to the time.

As per Fig. 7 which shows the comparison chart between various CNN algorithms such as Standard tensorflow algo and inception as well as faster r-CNN and ResNet, and it can be observed very clearly that Resnet 50 worked absolutely fine and with increased level accuracy.

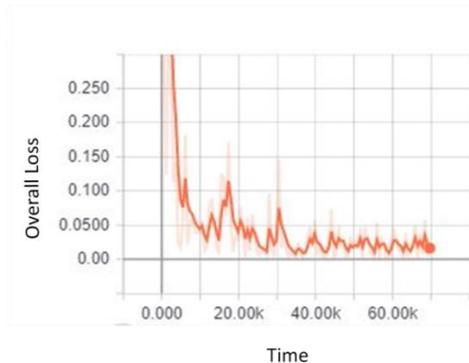


Fig. 6. Total loss during training

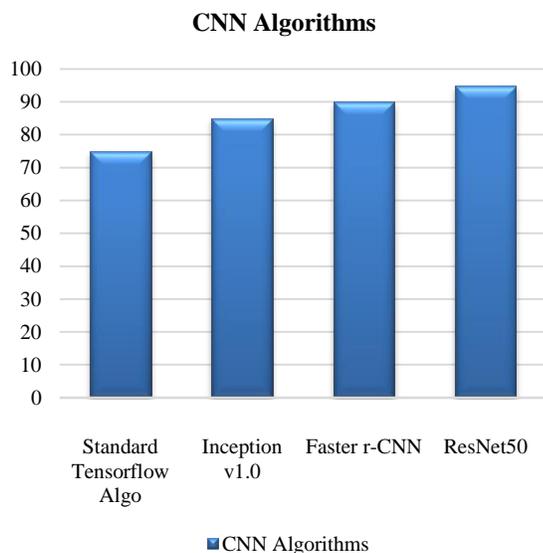


Fig. 7. CNN algorithms comparison chart

This paper proposes a live garbage detection product which can be used by private as well as public authorities alike. It can be developed further to provide a wide array of services like distinguishing between biodegradable, non-biodegradable and toxic wastes, classification of waste like glass, paper, polythene etc., even determining the best and most economical process for different types of waste segregated from the dumps.

## V. CONCLUSION

This research work proposes a solution to the garbage collection problem, faced by many in the modern communities. It is an automated garbage detection solution which can be used by garbage collection authorities as well societies alike. It entails live video analysis (like CCTVs) for identification of open garbage dumps and unofficial land dumps in streets, societies etc. For garbage detection several frameworks were utilized like the ResNet50 model and the Tensorflow technology to custom train a model. OpenCV3 along with other packages of python were implemented to run the live video feed. The implementation still needs to map every frame faster which can be done using better models and a larger dataset for training, so as to decrease the lag suffered during video testing. The dataset can also be implemented in a improved manner by specifying the dataset towards a state or city for better accuracy.

In the future versions of the project there are plans to implement-

- Classification of garbage into various categories like Biodegradable, Non-Biodegradable, Hazardous and Recyclable wastes.
- Train the model for a large dataset and differentiate waste into different classes like cardboard, glass, paper, plastic, tin/metals etc.
- Suggest the best methods to treat the various types of garbage identified by the software.
- Integrate the machine learning model and deploy it on mobile and various IOT devices.
- Shift the model on a cloud-based system for an even faster detection and better accuracy [15].

This venture can help in cleaning the streets of the country also increase public awareness over the subject as well. Some quality of life improvements are still needed in the project, but with the lack of other similar service in the market it's a step forward in the right direction.

This article mainly contribute towards identification and predict the garbage and dumps with the help of CCTV and video segments. The performance of the model is also at par from the other algorithms and comparatively satisfactory.

## ACKNOWLEDGMENT

This research was supported by Universiti Tun Hussein Onn Malaysia (UTHM) through Matching Grant (vot H995).

REFERENCES

- [1] Ying Liu and Zhishan Ge, "Research on automatic Garbage Detection System Based on Deep Learning and Narrowband Internet of Things", 2018.
- [2] Zhong-Qiu Zhao and Peng Zheng, "Object Detection with Deep Learning: A Review", 2017.
- [3] Huang, Jonathan & Rathod, Vivek & Sun, Chen & Zhu., "Speed/Accuracy Trade-Offs for Modern Convolutional Object Detectors", 2017
- [4] Nurminen, Jukka & Rainio, Kari & Numminen, Jukka-Pekka, "Object Detection in Design Diagrams with Machine Learning.", 2018.
- [5] Anitya & Kumar, Akhilesh & Bhushan, Vinayak , "World of intelligence defense object detection "machine learning", 2018.
- [6] Chen, Chunlin & Ling, Qiang," Adaptive Convolution for Object Detection", 2019.
- [7] Huang, Lili & Wu, Hefeng & Li, Guanbin & Wang, Qing, "Instruction-guided object detection.", 2019.
- [8] Kharinov, Mikhail & Buslavsky, A, "Object Detection in Color Image", 2019.
- [9] Liu, Ying & Ge, Zhishan & Lv, Guoyun & Wang, Shikai., "Research on Automatic Garbage Detection System Based on Deep Learning and Narrowband Internet of Things.", 2018.
- [10] Hu, Guo Qiang, Jing Chang Huang, Jun Chi Yan, and Jun Zhu. "Object detection." U.S. Patent 10,706,530, issued July 7, 2020.
- [11] Adrian Rosebrock, "Object detection with deep learning and OpenCV"
- [12] Lakshmi, V. Srinithi Santhana, et al. "Smart garbage alert system using machine learning." *Int. J. Eng. Appl. Sci. Technol* 5 (2020): 487-489.
- [13] Raccon Dataset, GitHub.com
- [14] Singh, Shubhendu, Kushal Samir Mehta, Nishant Bhattacharya, Jyotsna Prasad, S. Kaala Lakshmi, K. V. Subramaniam, and Dinkar Sitaram. "Identifying uncollected garbage in urban areas using crowdsourcing and machine learning." In 2017 IEEE Region 10 Symposium (TENSymp), pp. 1-5. IEEE, 2017.
- [15] Joshi, Jetendra, Joshitha Reddy, Praneeth Reddy, Akshay Agarwal, Rahul Agarwal, Amrit Bagga, and Abhinandan Bhargava. "Cloud computing based smart garbage monitoring system." In 2016 3rd International Conference on Electronic Design (ICED), pp. 70-75. IEEE, 2016.
- [16] Geiger, R. Stuart, Kevin Yu, Yanlai Yang, Mindy Dai, Jie Qiu, Rebekah Tang, and Jenny Huang. "Garbage in, garbage out? Do machine learning application papers in social computing report where human-labeled training data comes from?." In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, pp. 325-336. 2020.
- [17] Hoque, Mohammad Akidul, Mrittika Azad, and Md Ashik-Uz-Zaman. "IoT and Machine Learning Based Smart Garbage Management and Segregation Approach for Bangladesh." 2019 2nd International Conference on Innovation in Engineering and Technology . IEEE, 2019.
- [18] Shamin, N., P. Mohamed Fathimal, R. Raghavendran, and Kamallesh Prakash. "Smart garbage segregation & management system using Internet of Things (IoT) & Machine Learning (ML)." In 2019 1st International Conference on Innovations in Information and Communication Technology (ICIICT), pp. 1-6. IEEE, 2019..
- [19] Baby, Cyril Joe, Harvir Singh, Archit Srivastava, Ritwik Dhawan, and P. Mahalakshmi. "Smart bin: An intelligent waste alert and prediction system using machine learning approach." In 2017 international conference on wireless communications, signal processing and networking (WiSPNET), pp. 771-774. IEEE, 2017.
- [20] DeBrusk, Chris. "The risk of machine-learning bias (and how to prevent it)." MIT Sloan Management Review (2018).
- [21] Kim, In Kee, Sai Zeng, Christopher Young, Jinho Hwang, and Marty Humphrey. "iCSI: A cloud garbage VM collector for addressing inactive VMs with machine learning." In 2017 IEEE International Conference on Cloud Engineering (IC2E), pp. 17-28. IEEE, 2017.
- [22] Wang, Ying, and Xu Zhang. "Autonomous garbage detection for intelligent urban management." *MATEC Web of Conferences*. Vol. 232. EDP Sciences, 2018.
- [23] Park, Jung Kyu, and Jaeho Kim. "A method for reducing garbage collection overhead of SSD using machine learning algorithms." 2017 International Conference on Information and Communication Technology Convergence (ICTC). IEEE, 2017.
- [24] Gan, Ji Sheng. Automated Garbage Classification and Sorting System using Machine Learning. Diss. Tunku Abdul Rahman University College, 2020.
- [25] Sharma, Akhilesh K., Kamaljit I. Lakhtaria, Avinash Panwar, and Santosh Vishwakarma. "An analytical approach based on self organized maps (SOM) in Indian classical music raga clustering." In 2014 Seventh International Conference on Contemporary Computing (IC3), pp. 449-453. IEEE, 2014.
- [26] Sharma, Akhilesh K., Avinash Panwar, Prasun Chakrabarti, and Santosh Vishwakarma. "Categorization of ICMR Using feature extraction strategy and MIR with ensemble learning." *Procedia Computer Science* 57 (2015): 686-694.
- [27] Sharma, Akhilesh K., Avinash Panwar, and Prasun Chakrabarti. "Analytical approach on Indian classical raga measures by feature extraction with EM and Naive Bayes." *International Journal of Computer Applications* 107.6 (2014).
- [28] Sharma, Akhilesh K., and Prakash Ramani. "Rigorous data analysis and performance evaluation of Indian classical raga using RapidMiner." *Soft Computing: Theories and Applications*. Springer, Singapore, 2018. 97-106.
- [29] Sharma, Akhilesh Kumar, Gaurav Aggarwal, Sachit Bhardwaj, Prasun Chakrabarti, Tulika Chakrabarti, Jemal H. Abawajy, Siddhartha Bhattacharyya, Richa Mishra, Anirban Das, and Hairulnizam Mahdin. "Classification of Indian classical music with time-series matching deep learning approach." *IEEE Access* 9 (2021): 102041-102052.

# Tamper Proof Air Quality Management System using Blockchain

Vaneeta M<sup>1\*</sup>, Deepa S R<sup>2</sup>, Sangeetha V<sup>3</sup>, Kamalakshi Naganna<sup>4</sup>, Kruthika S Vasisht<sup>5</sup>, Ashwini J<sup>6</sup>, Nikitha M<sup>7</sup>,  
Srividya H.R<sup>8</sup>

Department of Artificial Intelligence and Machine Learning, K. S. Institute of Technology, Bangalore, India<sup>1\*</sup>

Department of Computer Science and Design, K. S. Institute of Technology, Bangalore, India<sup>2</sup>

Department of Computer Science and Engineering, Ramaiah Institute of Technology, Bangalore, India<sup>3</sup>

Department of Computer Science and Engineering, Sapthagiri College of Engineering, Bangalore, India<sup>4</sup>

Department of Computer Science and Engineering, K S Institute of Technology, Bangalore, India<sup>5</sup>

Department of Development, Gupshup Technologies, Bangalore, India<sup>6</sup>

Department of Development, TVAST IT Solutions, Bangalore, India<sup>7</sup>

Department of Development, Accolite Digital, Bangalore, India<sup>8</sup>

**Abstract**—One of the most important concerns facing urban regions across the world is air pollution. As a result, it's critical to monitor pollution levels and notify the public on the state of the air. An indicator called the Air Quality Index (AQI) does this by mapping the concentration of different contaminants into a single number. Because the examination of pollutant data is frequently opaque to outsiders, poor environmental control judgments may result. This has led to a need for a tamper-proof pollution management system for use by authorities, like the state and central pollution boards. To address these challenges, a model using machine learning algorithms to predict the air quality and store that information in the blockchain is proposed. Machine learning algorithms are used to categorize the air quality, and blockchain technology guarantees a permanent, tamper-proof record of all air quality data. Such a system might address the persistent issues with data dependability, immutability and trust in pollution control. The execution time of two main operations on blockchain are measured. The execution time of the put block is measured as 10 ms and the get block gets executed in 1 ms that fetches data from the blockchain ledger.

**Keywords**—Air pollution; air quality index; machine learning; blockchain technology

## I. INTRODUCTION

Air quality refers to how well the air is suited for breathing by people, animals, and plants. An average healthy person breathes approximately 14,000 liters of air each day. As a result, poor air quality may have an effect on the quality of life for both the present and future generations by hurting human well-being, the environment, the economy, and urban sustainability.

AQI Category (Range)	PM <sub>10</sub> (24hr)	PM <sub>2.5</sub> (24hr)	NO <sub>2</sub> (8hr)	O <sub>3</sub> (8hr)	CO (8hr)	SO <sub>2</sub> (24hr)	NH <sub>3</sub> (24hr)
Good (0-50)	0-50	0-30	0-40	0-50	0-1.0	0-40	0-200
Satisfactory (51-100)	51-100	31-60	41-80	51-100	1.1-2.0	41-80	201-400
Moderately polluted (101-200)	101-250	61-90	81-180	101-168	2.1-10	81-380	401-800
Poor (201-300)	251-350	91-120	181-280	169-208	10-17	381-800	801-1200
Very poor (301-400)	351-430	121-250	281-400	209-748	17-34	801-1600	1200-1800
Severe (401-500)	430+	250+	400+	748+	34+	1600+	1800+

Fig. 1. Air quality index (AQI) category range

The government keeps an eye on the air quality in various locations to determine the pollution level and to ensure that pollutant levels are within acceptable limits for human health. Air quality agencies can better plan how and when they will take action to safeguard the public's health by identifying how much pollution is present in a given location. Fig. 1 shows the AQI category range of the major pollutants.

The current technique for tracking industrial pollution is centralized, with a lack of openness and the possibility of data falsification. As a result, a consistent and tamper-proof mechanism must be utilized, such as secure software with data encryption and simultaneous data transfer directly to the regulator. Blockchain delivers Distributed ledger technology (DLT), which possesses the potential to solve many of the present system's open issues. Blockchain nodes are a network of multiple storage and computing devices that replicate data over a highly available and fault-tolerant infrastructure. Thus, blockchain facilitates the operation of a distributed database that is transparent and tamper-resistant. There is a need to design and develop an application using machine learning to predict Air quality category and store it on the blockchain that ensures it is tamper-proof and secure. The proposed system has three modules namely machine learning model, Blockchain network and Client application

The machine learning model is trained using industrial air pollution data. Supervised learning algorithms such as random forest classifier, decision tree classifier and Naive Bayes are used to predict the air quality index and the quality range of the given input data. The design of the ML model has these phases. The dataset comprises pollutant concentration information from over 15 industrial areas across India. This data set has around 37-40 pollutants, but the seven most appropriate pollutants are considered. The dataset is cleaned and partitioned into training and testing data. On comparing the results, best results were obtained from decision tree classifier with an accuracy of 99.6%.

The next module is the Blockchain network. The chaincode contains the ML model deployed in it. Once the client supplies the data to the blockchain, the chaincode that has the ML

model will start executing. Once the air quality category has been predicted, it is sent as a transaction to the users of blockchain for endorsement according to the endorsement policy decided prior. The users of this system are the members of state and central pollution control boards. The transaction is then ordered and validated by the respective chain code. If the transaction is validated, it is stored on the ledger.

S. Benedict [1] proposed a blockchain enabled IoT cloud implementation to tackle the existing issues of security hazards and performance inefficiencies in smart cities. It particularly highlights the implementation of chaincodes for air quality monitoring systems in Smart Cities. The proposed architecture named as IoT enabled Blockchain for Air Quality Monitoring System (IB-AQMS) is illustrated using experiments.

Abu Buker [2] proposed an indoor AQI monitoring system to predict the AQI through the Neural Network Algorithm and Block-chain. The Indoor Air Quality system consists of sensors such as temperature, humidity, Carbon Di Oxide, Particulate Matter, Carbon Mono Oxide, and LPG. The Neural Network decision-making model is used to predict the AQI. The suggested IoT-based smart block-chain technology plays a vital role by imparting scalability, privacy, and reliability.

The significant objectives of proposed work are:

- To design and experiment with Random Forest classifier, Decision tree classifier and Linear regression algorithms to predict air quality category and consider the one with best accuracy.
- To implement blockchain-based solutions to resolve the ongoing issues with data dependability in pollution monitoring.
- To ensure a permanent, immutable record of all air quality data of industries.
- To develop GUI for the user to maintain an immutable record of all air quality information.

The organization of the paper is as follows: Section II discusses about the background literatures used for modelling AQI index and use of blockchain based technology for storing data. Section III elaborates the proposed solution and discusses about detailed implementation followed by discussion of result in Section IV. Finally, the conclusive remarks are provided in Section V and future scope is presented in Section VI.

## II. BACKGROUND

The work [3] by S. Mahanta investigates the efficacy of different existing prediction models in forecasting AQI values based on input values. According to Dyuthi Sanjeev's article [4], the AQI is calculated based on pollutants or attributes that have the greatest impact on air pollution. The Random Forest model is the most efficient, according to the research, with a score of 99.4 percent accuracy. Timothy M's research [5] proposes a method for evaluating air quality by building prediction models that link sensor data to an air quality score. Aditya C R [6] uses logistic regression to determine if a data sample is contaminated, as well as auto regression to predict future PM2.5 values based on present PM2.5 data. The study's

purpose [7] is to examine a range of existing prediction models to see how effective they are at predicting data from the study area.

Yue-Shan Changa [8] offers an ALSTM (Aggregated Long Short-Term Memory Model) that combines regional air-quality monitoring stations, industrial zone stations, and external emission source stations. Mahmoud Reza Delavar's [9] study provides a novel method for predicting air pollution in urban regions based on both stationary and non-stationary sources, using machine learning and statistical approaches.

This study [10] will be using Data Mining and Machine Learning models in this research project, to forecast the AQI and classify the AQI into buckets labeled as Good, Satisfactory, Moderate, Poor, Very Poor and Severe. Regression models are used to predict AQI. In order to predict AQI bucket, KNN (K Nearest Neighbors) algorithm with repeat CV classification is used. Station-level data from Indian cities was used to accurately classify and forecast AQI Labels. KNN and repeat CV classification performed best in terms of accuracy. M. Lücking et al. [11] offer a software design for a pollution monitoring system (PMS) based on distributed ledger technology and the long-range protocol, which is described in this paper. It provides adaptable, traceable, and energy-efficient monitoring. Multiple unresolved issues in the functioning of pollution monitoring systems, such as storing data that is invalid or susceptible to tampering, are addressed by distributed ledger technology in a Hyperledger Fabric blockchain.

One of the prime components of blockchain is cryptography for providing confidentiality and authentication using efficient keys. In regard to this author Vaneeta in paper [12] proposes multi-tier framework by including a superior authentication scheme using enhanced public key encryption and digital signature. Sina Rafati Niya's paper [13] proposes an automated approach for measuring, monitoring, and storing air and water quality in factories, lakes, and other sites, based on an IoT and Blockchain-based system.

The proposed system in paper [14] collects real-time air pollution data from industrial locations using 5G wireless IoT sensors and transmits encrypted blockchain data to the index measurement service and cloud via a periodic blockchain transaction. This device enables real-time pollution monitoring in industrial settings and also protects data from tampering. The distributed messaging protocol and blockchain's encryption technologies increase the efficiency of data processing and exchange, while maintaining data integrity. The main objective of this research study [15] is to give an overview of technologies such as Artificial Intelligence, Blockchain and Internet of Things (IoT) and their current applications in the fields of public healthcare and the environment.

Air pollution has been a source of great concern for a long time, but it has come to the attention of stakeholders only recently. The Air Act of 1981 was the piece of legislation that established the requirement for air quality monitoring and opened the door to the monitoring techniques employed in India under the CPCB's oversight. This study [16] suggests that we adopt a thorough approach to manage air pollution. In this

paper [17] development of less expensive, simple-to-use, portable air pollution monitoring sensors, which deliver high-time resolution data in almost real-time and makes access to environmental data convenient is discussed. A variety of air contaminants can already be monitored by sensor devices, and new technologies are constantly being developed.

Regulating and monitoring pollution emissions becomes increasingly essential for battling the illness. This research [18] proposes an Internet of Things (IoT) based system that uses low sensors to monitor pollutants. It is developing a hardware layer of device that is capable of measuring concentrations of pollutants by means of three sensors, respectively, MQ-131, PMSA003 and MICS-6814. The given research study [19] proposes use of Internet of Things (IoT) sensors to periodically collect air quality information such as pollutant concentration and transmit the same over Low power wide area (LPWA) network. IoT cloud is used to process and analyze the data. The participatory urban sensing architecture for PM<sub>2.5</sub> monitoring described in this research [20] has more than 2500 devices operating in Taiwan and 29 other nations. The framework's open system design, which is built on the ideas of open hardware, open source software, and open data, is its defining feature.

The paper [21] introduces CNN-ILSTM, an Air Quality Index prediction model based on Convolution Neural Networks (CNN) and Improved Long Short-Term Memory (ILSTM). The experimental data set includes air quality data from 00:00 on April 4, 2019 to 23:00 on June 30, 2021 in Shijiazhuang, Hebei Province, China. Air pollution is affecting public health and causing a slew of health-related issues, resulting in a significant medical bill each year. Taking air quality information into account, the study [22] offers a safe path when the air quality index is poor to reduce the impact on human health. Dijkstra's method is used to discover the safest path between source and destination. This study also points out research gaps in various studies on similar grounds. The first completely expandable blockchain architecture for supporting distributed applications is called Hyperledger Fabric. Additionally, Fabric is the first blockchain platform to support distributed applications created in common, all-purpose programming languages, independent of a native coin as a system backend.

As indicated in the study [23], Fabric proposes a revolutionary architecture that is evocative of middleware-replicated databases and isolates transaction execution from consensus while enabling policy-based endorsement. Author Sangeetha in paper [24] proposed that Security issues are one of the core problems in mobile adhoc networks owing to the decentralized architecture. The proposed system introduces a new scheme that acts as multi-layer security under two different stages and enhances security in MANET's by modelling the different interactions among a malicious node and with a number of legitimate nodes.

Based on two publicly accessible datasets, this paper [25] regression models using support vector regression (SVR) and random forest regression (RFR) to predict the Air Quality Index (AQI) in Beijing and the nitrogen oxides (NOX) content in an Italian city. The performance of the regression models was assessed using the root-mean-square error (RMSE), correlation coefficient (r), and coefficient of determination (R<sup>2</sup>). SVR based model predicts AQI better whereas RFR based model predicts NOX concentration better. According to study in this paper [26], most of the research on Air quality uses Machine learning techniques and Big data analytics on data collected by IoT sensors. Aim of this study is to evaluate such techniques on air quality forecasting. Based on the observations made, study suggests the need for more research and development in real time air quality monitoring. To suffice the needs of future cities, an integrated air quality monitor with hybrid machine learning models can be developed that address impacts of dynamic quality on various atmospheric levels.

Research paper [27] proposes a real-time IoT based system for air quality monitoring. Study uses models such as CNN-LSTM-BOA (Convolution Neural Networks-Long Short Term Memory-Bayesian Optimization Algorithm) and other baseline models such as SVM (Support Vector Machine), ANN (Artificial Neural Network), Ensemble model. The LSTM model proved to be good for prediction. Results suggest that the proposed models perform better than baseline models. As part of further research, this study also suggests using statistical criteria, AI algorithms can evaluate performance and compare the results to publicly available data sets. Proposed technique in paper [28] used IoT sensors and Artificial Intelligence techniques that are said to reduce implementation costs to ⅓ since before. The proposed model includes wireless sensor nodes to measure gas concentrations (MQ series) which are connected by IEEE 802.11 Wireless LAN AP to an IoT cloud that stores and maintains data in turn connected to a machine learning model responsible for predicting air quality levels. Model uses ARIMA prediction technology which stands for differential autoregressive moving normal model.

Research [29] proposes a predictive model using a multilayer perceptron, support vector regression and linear regression to predict future condition of air quality in a vehicle, based on data collected from sensors. Performance of these models can be evaluated using Root mean square error, coefficient of determination (R<sup>2</sup>), Mean Squared Error and Mean Absolute Error. For data collected, the support vector regression model had the highest performance in terms of R<sup>2</sup> and had a lower error rate.

### III. PROPOSED SOLUTION

The suggested system's mechanism is depicted in Fig. 2. It demonstrates how the entire technique is broken down into five system implementation components.

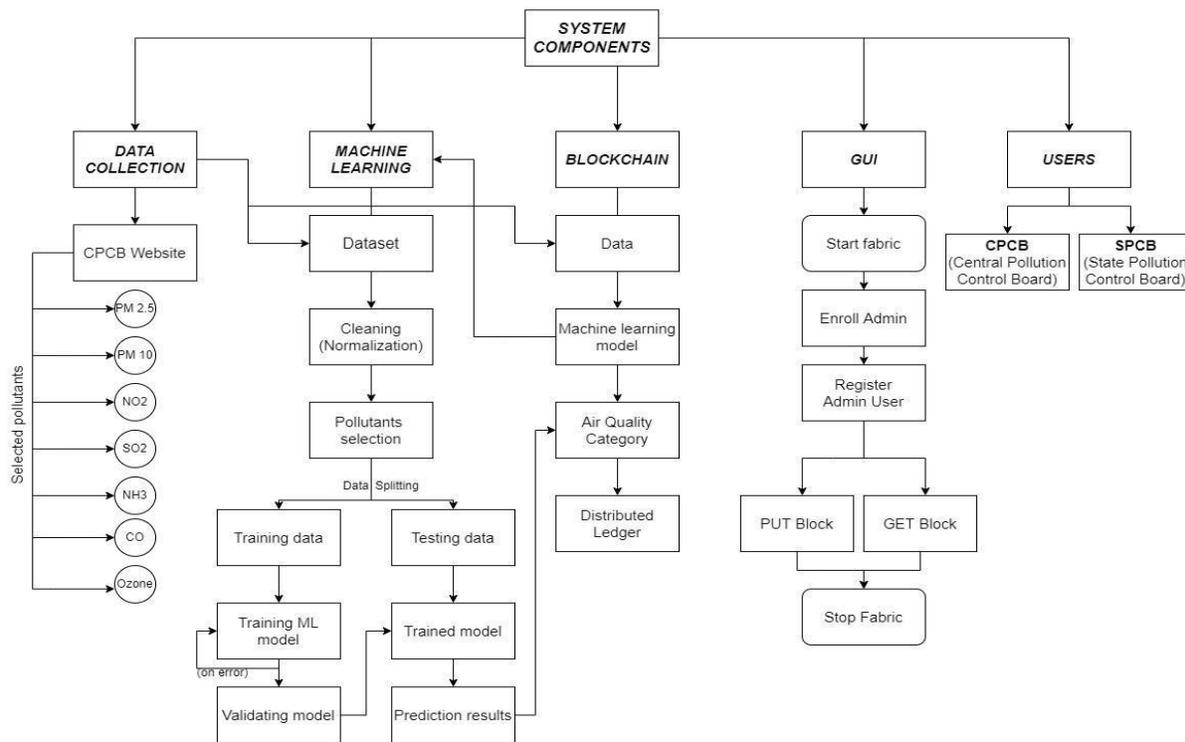


Fig. 2. System components

#### A. Data Collection

- Web scraping is used to collect data for the seven pollutants directly from the official CPCB website. This scraped data is given as input into the machine learning model deployed as the chaincode on the blockchain network. Chaincode then gives the AQI category as Good, Satisfactory, Moderately polluted, Poor, Very Poor and severe. The values of all seven pollutants and AQI categories are then stored on the distributed ledger.
- The web scraping process is divided into three steps. The first step is to set up to and from dates. For these consecutive dates 24hr-data is received. Second, is to pull the data for the setup dates in JSON format and finally, to parse this data and store this in a table form. The data is collected from an industrial area named Peenya in the city of Bangalore in Karnataka, India which is one of the biggest industrial areas in Asia.
- The data set is obtained from the Central Pollution Control Board's official website (CPCB). The website has around 15 stations in industrial areas all over India. The data is from Jan 1st 2020 till April 23rd 2022 for seven pollutants. Data is being considered on a 24-hour basis for five pollutants (PM2.5, PM10, NO2, SO2, NH3) and eight hourly basis for two pollutants (CO and ozone). The dataset has around 12600 rows.

#### B. Machine Learning Model

- The above air pollution data set is divided 80 percent for the machine learning model's training and 20 percent for its testing. Supervised learning methods-Linear regression, Random Forest classifier, and

Decision Tree classifier were the three machine learning techniques used to predict the air quality category of the given input data. The core nodes of the decision tree classifier represent dataset properties, the branches represent decision rules, and each leaf node represents the classification outcome. A 99.56 percent accuracy rate was generated by the decision tree classifier.

- Random forest algorithm is considered as an ensemble learning algorithm. The algorithm's core concept is to construct short and weak decision-trees with few attributes, in parallel, and then merge the trees to generate a single, powerful learning model by taking the majority vote or by just taking the average. An accuracy score of 99.06 % was observed for this model.
- Linear regression, a supervised learning algorithm, is the most common regression model which is used to determine how the independent variable(s) and the dependent variable are related. It is employed to ascertain how the value of the independent variable affects the value of the dependent variable. Finding the best fit line is essential when using linear regression since it reduces the error between the actual and projected values. A sloping straight line is used to represent the connection between the variables. The line that fits the data the best is the one with the least error or inaccuracy. An accuracy score of 91.79 % was observed for this model. Comparing the three learning algorithms. Decision tree classifier produced the most accurate results, with a 99.56 percent accuracy.

### C. Blockchain Network

A blockchain is a type of distributed ledger or database that securely and impenetrably keeps a chain of data in the form of blocks in chronological order. The chain of blocks, also known as a ledger, is continuously expanding, thus new blocks are added to the end of the ledger. Each new block retains a reference to the content of the preceding block using a hash value. The distributed ledger material is secured using the public key encryption process, which also ensures consistency, irreversibility, and non-reputability. The block's immutability, anonymity, and compactness are guaranteed by the use of a cryptographic one-way hash function, such as SHA256. In a peer-to-peer (P2P) network, the ledger and its contents are copied and synced among several peers, forming a distributed ledger. There are three basic categories of blockchains: consortium blockchains, private permissioned blockchains, and public permissionless blockchains. All blockchain data is open to and visible to the general public since the permissionless blockchain type stresses the public component. The Bitcoin and Ethereum blockchains are two examples of such a blockchain. A private blockchain, on the other hand, permits only selected nodes to join the network, making it appear to be a type of distributed but nonetheless centralized network. Combining the two, the consortium blockchain only allows a predetermined set of nodes to take part in the distributed consensus process.

The proposed work implements private blockchain using Hyperledger Fabric. Hyperledger Fabric is a distributed ledger platform that is open-source and enterprise-ready. It features extensive privacy controls that ensure that only the information you want shared with "permissioned" (known) network participants is shared.

The operation of fault-tolerant distributed ledgers is principally due to distributed ledger technology (DLT). Each distributed ledger node keeps a local copy of the data, and new data is added to the ledger in the form of transactions. New transactions are validated using digital signatures and saved in node memory, which is then passed on to other DLT nodes in the network. Eventually, approved transactions are either directly put to the ledger or recorded in a block and then added to the ledger. Most DLT consensus systems (like Kafka or Raft) or even Byzantine fault-tolerant ones are crash-tolerant (e.g., Nakamoto consensus). Crash fault tolerance is the capacity of a consensus mechanism to reach consensus across all validating nodes notwithstanding the (temporary) unavailability of nodes.

Comparing private-permissioned distributed ledgers versus public-permissionless distributed ledgers, the former frequently provides greater flexibility (maintainability), faster speed (rapid transaction confirmation and maximum throughput), and a high level of transparency.

The users of our blockchain network are the State Pollution Control Boards (SPCB), Central Pollution Control Board organization (CPCB) and an orderer organization. All the three organizations form a private "subnet" of communication called channel as shown in Fig. 3.

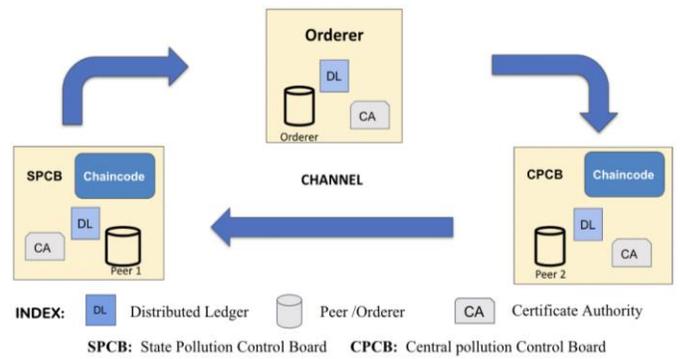


Fig. 3. Blockchain architecture

Chaincode (smart contracts) in Hyperledger Fabric are small programs written in Go, JavaScript/TypeScript or Java that contain the business logic to be executed as transactions on the blockchain. The chaincode has methods to store data received from machine learning models and query the ledger.

To digitally sign the response and endorse the transaction, the endorsement system chaincode is employed. After the transaction is ordered, a validation system chaincode compares the endorsements in the transaction to the endorsement policy defined for the chaincode. If the policy on endorsement isn't followed, then that transaction is marked invalid. Once the transaction has been endorsed, it is ordered and validated by respective peers and chaincode. If the transaction is validated successfully, it is stored on the ledger and every peer will maintain a copy of the same. Otherwise it is rendered unsuccessful.

Every transaction should be endorsed by either peers of SPCB or CPCB organizations, which is mentioned in the endorsement policy. Orderer nodes sequence groups of approved transactions into blocks and bundle them. These blocks are added to the blockchain. The Orderer then distributes blocks to all peers associated with it. Every peer validates the distributed block separately, maintaining consistency, to ensure the block is endorsed by the peers of the right organization and follows the endorsement policy. To prevent altering the ledger's state, all invalidated blocks are appended to an immutable block that was produced by the orderer and was designated invalid by the peer.

For a transaction to occur between the client application and the blockchain network, the client application primarily needs to have a certificate saying it can interact with the blockchain network and should have the necessary information of the network. The steps in transaction are as shown in Fig. 4.

1) The client application enrolls a user in order to get a valid certificate required to communicate with the Hyperledger Fabric blockchain network. Next the client application calls one or more peers to discover the network and since Hyperledger fabric is a private permission blockchain, users get to access only a part of the network. This step is not to be performed before every transaction. Once the required certificate along with topology of the required part of the blockchain network is obtained by the client application, the actual transaction process can be started.

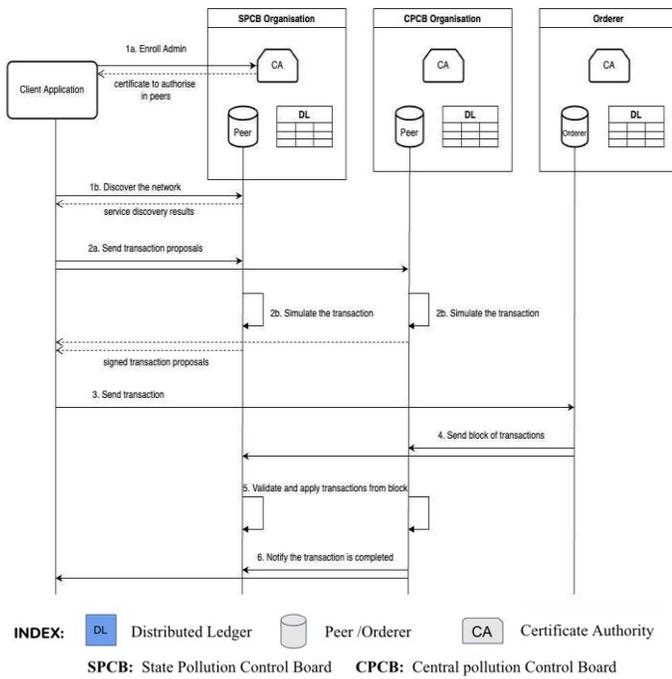


Fig. 4. Transaction proposal

2) Transaction proposals are sent to peers by the client application. These transactions should satisfy the endorsement policy, hence peers simulate the transaction by calling the smart contract which determines what has to be read and what has to be written to the world state based on their copy of the ledger if the transaction succeeds. This information along with the digital signatures from the peers is returned to the client application.

3) The client application next sends the transaction (which contains the simulation results and peer signatures) to the ordering service.

4) The ordering service creates a block once it has collected, validated, and ordered the appropriate number of transactions. The block is then transmitted to the channel's lead peers, who pass it along to the other peers.

5) The transactions are validated and applied by each peer that receives the block. The world state databases are updated with the transaction read/write sets as well as the blockchain copies on the peers are appended with new transactions.

6) The client application is expected to wait until relevant peers notify them that the transaction has been successfully completed. This notification indicates that it actually was appended to the blockchain network on a given peer.

Fig. 5 presents the flowchart of GUI system. The user application is for SPCB and CPCB. Once they access the website they can click on "START NETWORK" to start the network setup. Once the network is up and running, one can enroll the admin by clicking "ENROLL ADMIN" and once this is done, the admin user can be registered through "REGISTER USER".

#### D. GUI

The major functionality of the client is to collect values of pollutants on a 24 hr basis and transfer these to the blockchain along with the details of the concerned industrial area. The user inputs the "to" and "from" date after clicking the "PUT BLOCK" button as shown in Fig. 5 for which the required data is to be extracted. Once the date is entered, the concentration values of each pollutant for that particular "to"-date is fetched and displayed with the help of an API designed by us. The pollution data is then sent to the ML model with the help of another API. The model then calculates the AQI and determines the AQI category. This predicted category is then stored in the blockchain ledger along with the prefetched pollution data. Further, the ledger can be accessed through the GUI by clicking on the "GET BLOCK" button and the contents are displayed as shown in Fig. 5.

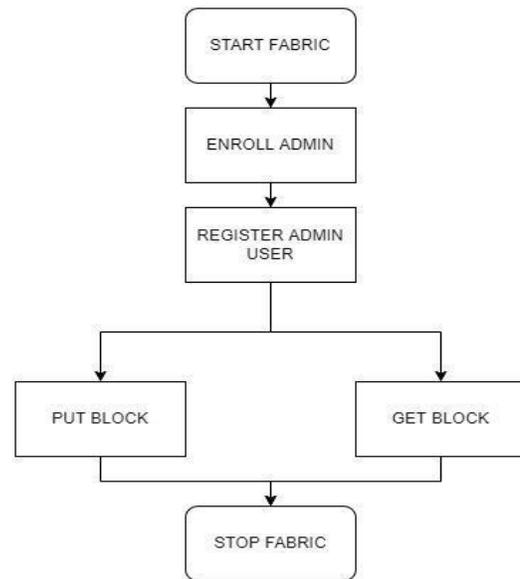


Fig. 5. GUI of the system

#### E. Users

The users of our system are the State pollution control board (SPCB) and the Central pollution control board (CPCB).

The SPCB is a council that analyzes, supervises, and partakes in an inquiry. The Board has a team of experts and a testing facility to evaluate the quality of various samples taken from industrial areas' soil, water, and air samples. It operates in accordance with the guidelines that the government occasionally sets. The CPCB advises the government and SPCB on issues pertaining to the implementation and enforcement of the Air, Water, and Environmental Acts.

#### IV. SYSTEM IMPLEMENTATION

Steps in setting up blockchain Network:

1) *Generate certificates using cryptogen tool:* All entities must be first recognized and granted permission before entering a consortium network under a permissioned blockchain. With the help of a bin/cryptogen tool provided by Hyperledger Fabric, crypto material is created. The

cryptographic components of the Test Network are created using a configuration file, and the finished product is stored in the directory structure. Along with these and docker compose files consortium network can be started.

2) *Generating orderer genesis block*: A "genesis block" is the initial block of a freshly established channel and first block of the "orderer system channel".

3) *Generating channel configuration transaction 'channel.tx'*: The shared configuration for a Hyperledger Fabric blockchain network is stored in a set of configuration transactions, one per channel.

4) *Generating anchor peer update for Org1MSP and Org2MSP*: Peers that are present outside of the organization are known as anchor peers. Any communication between organizations needs anchor peers.

Build your first network (BYFN) end-to-end test

Channel name: mychannel

Creating a channel...

Channel 'mychannel' created

This will launch all of the containers, and then drive a complete end-to-end application scenario.

5) *Having all peers join the channel...*

- a) peer1.org1 joined channel 'mychannel'
- b) peer0.org2 joined channel 'mychannel'
- c) peer1.org2 joined channel 'mychannel'

Administrators can conduct channel-related actions on a peer, such as joining the peer to a channel, using the peer channel command.

6) *Updating anchor peers for org1...*

a) Anchor peers updated for org 'Org1MSP' on channel 'mychannel'

Updating anchor peers for org2...

b) Anchor peers updated for org 'Org2MSP' on channel 'mychannel'

7) *All GOOD, BYFN execution completed*

- Installing smart contract on peer0.org1.example.com
- Installing smart contract on peer0.org2.example.com
- Instantiating smart contract on mychannel
- Waiting for instantiation request to be committed ...
- Submitting initLedger transaction to smart contract on mychannel
- The transaction is sent to the two peers with the chaincode installed (peer0.org1.example.com and peer0.org2.example.com) so that chaincode is built before receiving the following requests

8) *Total setup execution time: 181 secs ...*

## V. RESULTS AND DISCUSSION

The below Fig. 6 shows the User Interface with start network, Enroll admin and register user button options. Once the start network option is selected the blockchain network is established between two organizations of CPCB and SPCB. During start network steps 1 to 8 get executed. After the network is up the admin is enrolled by entering user name and password. Once the admin is logged the user can register by entering username and password.

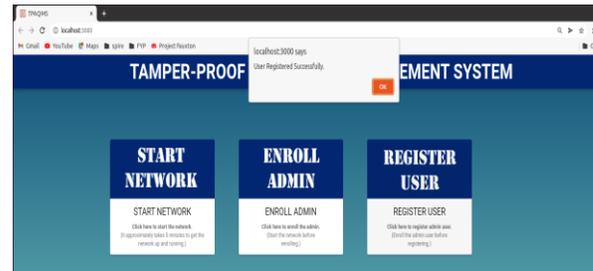


Fig. 6. User Interface

The Fig. 6 shows the User Interface for inserting the from and to dates for which the data must be fetched from the CPCB website. After fetching data of pollutants from CPCB website the Machine Learning model for calculation of AQI index is executed and its air quality category is identified. Put Block option places the transaction consisting of pollutant values along with AQI category on the blockchain ledger. Get block option retrieves block from ledger. The blockchain network can be stopped by selecting Stop Network option (see Fig. 7).

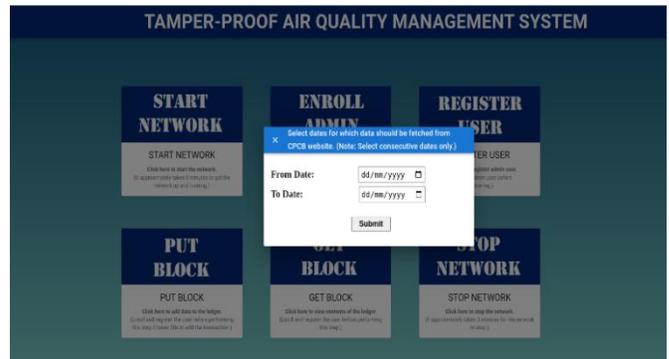


Fig. 7. Input dates to get pollutants

Fig. 8 shows the ledger data maintained by blockchain network accessible only to two organizations. It stores date with pollutant values and AQI category.

Date	PM2.5	PM10	NO2	SO2	NO	CO	Count	AQI Category
10 Jun 2022	37.51	93.24	8.99	16.93	36.86	0.57	2812	Moderately Polluted
15 Jun 2022	31.44	67.55	10.2	17.17	33.31	0.52	2937	Moderately Polluted
19 Jun 2022	27.9	72.35	13.97	9.05	24.59	0.5	414	Satisfactory

Fig. 8. Ledger data

The performance parameter considered is time required for setting the blockchain network. The graph in Fig. 9 represents the average execution time of certain transactions in blockchain network.

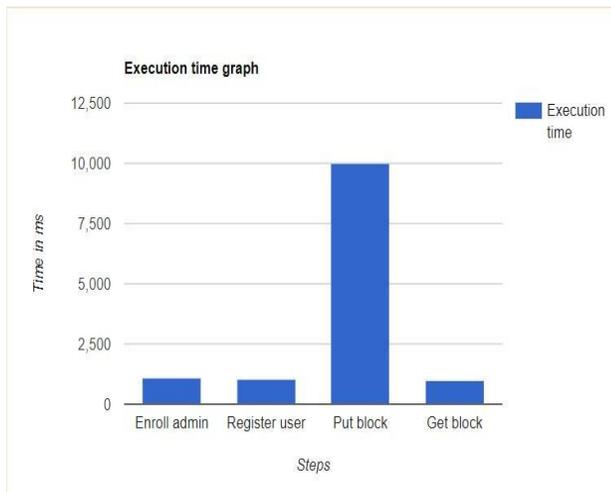


Fig. 9. Execution time

The configuration of a computer system considered is 64bit OS, 12 GB RAM, i7 core processor, 1.99 GHz. For a transaction to occur in a private blockchain network such as Hyperledger fabric, the client application needs to enroll and register the user whose execution times are 1000ms. The execution time of the put block is 10,000ms which includes fetching pollution data blockchain ledger. Lastly, the get block gets executed in 1000ms that fetches data from the blockchain ledger from the CPCB website that is sent to the ML model which determines the AQI category that is to be stored in the blockchain ledger.

The proposed system uses Hyperledger Fabric blockchain network. It is a private network. Hyperledger Fabric is private, permissioned network and does not use currencies. Proposed system is a simple attempt to store transaction securely on ledger. Ethereum blockchain network used in [30] is private or public without any permissions for users and even uses currency called as ether. As a solution for verifying the accuracy of sensitive pollution data, offers a blockchain management system. Data provided by an air quality monitoring network with high geographical and temporal resolution could be traced back in time. Ethereum blockchain supports preserving data on the average densities of zones regarded to be the city's key areas.

## VI. CONCLUSION

The present system for tracking pollutants emitted in industrial areas is centralized, lacks complete transparency and is highly susceptible to data tampering. Proposed tamper-proof air quality management platform combines a machine learning model to predict the AQI category which is then stored on Blockchain. Prediction accuracy has improved using the Decision tree algorithm which is a machine learning algorithm that gives us an accuracy of 99.56%.

It is discovered that a blockchain-based solution can address data dependability issues in pollution monitoring while

also providing a permanent, tamper-proof record of industrial air quality data. As a result, industrial area-specific air quality data may be provided in a credible and transparent manner, allowing industries and the government to take the required steps to minimize pollution.

## VI. FUTURE SCOPE

The proposed tamper proof air quality management system is limited to collecting pollutant data from CPCB website calculating AQI category and storing on blockchain ledger of two organizations. The system can be enhanced by incorporating Internet of Things for collection for pollutant values. IoT device with sensors for different pollutants can be installed and real time data can be collected, calculate the AQI category and store the transaction on blockchain ledger.

## REFERENCES

- [1] S. Benedict, P. Rumaise and J. Kaur, "IoT Blockchain Solution for Air Quality Monitoring in SmartCities," 2019 IEEE International Conference on Advanced Networks and Telecommunications Systems (ANTS), Goa, India, 2019, pp. 1-6, doi: 10.1109/ANTS47819.2019.9118148.
- [2] Abu Buker Siddique, Rafaqat Kazmi, Habib Ullah Khan, Sikandar Ali, Ali Samad & Gulraiz Javaid (2022) An Intelligent and Secure Air Quality Monitoring System Using Neural Network Algorithm and Blockchain, IETE Journal of Research, DOI: 10.1080/03772063.2022.2052984
- [3] S. Mahanta, T. Ramakrishnu, R. R. Jha and N. Tailor, "Urban Air Quality Prediction Using Regression Analysis," TENCON 2019 - 2019 IEEE Region 10 Conference (TENCON), 2019, pp. 1118-1123, doi: 10.1109/TENCON.2019.8929517.
- [4] Dyuthi Sanjeev, 2021, Implementation of Machine Learning Algorithms for Analysis and Prediction of Air Quality, INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT) Volume 10, Issue 03 (March 2021)
- [5] T. M. Amado and J. C. Dela Cruz, "Development of Machine Learning-based Predictive Models for Air Quality Monitoring and Characterization," TENCON 2018 - 2018 IEEE Region 10 Conference, 2018, pp. 0668-0672, doi: 10.1109/TENCON.2018.8650518.
- [6] C R, Aditya & Deshmukh, Chandana & K, Nayana & Gandhi, Praveen & astu, Vidyav. (2018). Detection and Prediction of Air Pollution using Machine Learning Models. International Journal of Engineering Trends and Technology. 59. 204-207. 10.14445/22315381/IJETT-V59P238.
- [7] Stojov, Vladimir & Koteli, Nikola & Lameski, Petre & Zdravetski, Eftim. (2018). Application of machine learning and time-series analysis for air pollution prediction.
- [8] Yue-Shan Chang, Hsin-Ta Chiao, Satheesh Abimannan, Yo-Ping Huang, Yi-Ting Tsai, Kuan-Ming Lin, An LSTM-based aggregated model for air pollution forecasting, Atmospheric Pollution Research, Volume 11, Issue 8, 2020.
- [9] Delavar, M.R.; Gholami, A.; Shiran, G.R.; Rashidi, Y.; Nakhaeizadeh, G.R.; Fedra, K.; Hatefi Afshar, S. A Novel Method for Improving Air Pollution Prediction Based on Machine Learning Approaches: A Case Study Applied to the Capital City of Tehran. ISPRS Int. J. Geo-Inf. 2019, 8, 99.
- [10] Londhe, Mayuresh. "Data mining and machine learning approach for air quality index prediction." International Journal of Engineering and Applied Physics 1.2 (2021): 136-153.
- [11] M. Lücking et al., "The Merits of a Decentralized Pollution-Monitoring System Based on Distributed Ledger Technology," in IEEE Access, vol. 8, pp. 189365-189381, 2020, doi: 10.1109/ACCESS.2020.3028430.
- [12] Vaneta M, S Swapna Kumar, "Multi-tier Framework for Optimizing Pairwise Key Predistribution in Sensory Applications", International Journal of Innovative Technology and Exploring Engineering, 2019, 2278-3075, Volume-9, Issue-2.

- [13] S. R. Niya, S. S. Jha, T. Bocek and B. Stiller, "Design and implementation of an automated and decentralized pollution monitoring system with blockchains, smart contracts, and LoRaWAN," NOMS 2018 - 2018 IEEE/IFIP Network Operations and Management Symposium, 2018, pp. 1-4, doi: 10.1109/NOMS.2018.8406329.
- [14] Yohan Han, Byungjun Park, Jongpil Jeong, A Novel Architecture of Air Pollution Measurement Platform Using 5G and Blockchain for Industrial IoT Applications, *Procedia Computer Science*, Volume 155, 2019, Pages 728-733, ISSN 1877-0509 <https://doi.org/10.1016/j.procs.2019.08.105>.
- [15] M. Bublitz, F.; Oetomo, A.; S. Sahu, K.; Kuang, A.; X. Fadrigue, L.; E. Velmovitsky, P.; M. Nobrega, R.; P. Morita, P. Disruptive Technologies for Environment and Health Research: An Overview of Artificial Intelligence, Blockchain, and Internet of Things. *Int. J. Environ. Res. Public Health* 2019, 16, 3847. <https://doi.org/10.3390/ijerph16203847>
- [16] Sunil Gulia, Nidhi Shukla, Lavanya Padhi, Parthaa Bosu, S.K. Goyal, Rakesh Kumar, Evolution of air pollution management policies and related research in India, *Environmental Challenges*, Volume 6, 2022, 100431, ISSN 2667-0100, <https://doi.org/10.1016/j.envc.2021.100431>. (<https://www.sciencedirect.com/science/article/pii/S2667010021004054>)
- [17] The Changing Paradigm of Air Pollution Monitoring Emily G. Snyder, Timothy H. Watkins, Paul A. Solomon, Eben D. Thoma, Ronald W. Williams, Gayle S. W. Hagler, David Shelow, David A. Hindin, Vasu J. Kilaru, and Peter W. Preuss *Environmental Science & Technology* 2013 47 (20), 11369-11377 DOI: 10.1021/es4022602
- [18] H. P. L. d. Medeiros and G. Girão, "An IoT-based Air Quality Monitoring Platform," 2020 IEEE International Smart Cities Conference (ISC2), 2020, pp. 1-6, doi: 10.1109/ISC251055.2020.9239070
- [19] K. Zheng, S. Zhao, Z. Yang, X. Xiong and W. Xiang, "Design and Implementation of LPWA-Based Air Quality Monitoring System," in *IEEE Access*, vol. 4, pp. 3238-3245, 2016, doi: 10.1109/ACCESS.2016.2582153.
- [20] L. -J. Chen et al., "An Open Framework for Participatory PM2.5 Monitoring in Smart Cities," in *IEEE Access*, vol. 5, pp. 14441-14454, 2017, doi: 10.1109/ACCESS.2017.2723919.
- [21] Wang J, Li X, Jin L, Li J, Sun Q, Wang H. An air quality index prediction model based on CNN-ILSTM. *Sci Rep.* 2022 May 19;12(1):8373. doi: 10.1038/s41598-022-12355-6. PMID: 35589914; PMCID: PMC9120089.
- [22] Shriram, Pranav and Malladi, Srinivas, A Study and Analysis of Air Quality Index and Related Health Impact on Public Health (January 18, 2021). ICICNIS 2020, Available at SSRN: <https://ssrn.com/abstract=3768477>
- [23] Androulaki, E., "Hyperledger Fabric: A Distributed Operating System for Permissioned Blockchains"
- [24] Sangeetha V, Vaneeta M, S Swapna Kumar, Piyush Kumar Pareek, Sunanda Dixit, "Efficient Intrusion detection of malicious node using Bayesian Hybrid Detection in MANET", CCRDA 2020, IOP Conf. Series: Materials Science and Engineering 1022 (2021) 012077, IOP Publishing, doi:10.1088/1757-899X/1022/1/012077,2021.
- [25] Liu, H.; Li, Q.; Yu, D.; Gu, Y. Air Quality Index and Air Pollutant Concentration Prediction Based on Machine Learning Algorithms. *Appl. Sci.* 2019, 9, 4069. <https://doi.org/10.3390/app9194069>
- [26] Kaur, Gaganjot & Gao, Jerry & Chiao, Sen & Lu, Shengqiang & Xie, Gang. (2018). Air Quality Prediction: Big Data and Machine Learning Approaches. *International Journal of Environmental Science and Development*. 9. 8-16. 10.18178/ijesd.2018.9.1.1066.
- [27] Kataria, A., Puri, V.: AI- and IoT-based hybrid model for air quality prediction in a smart city with network assistance. *IET Netw.* 11( 6), 221– 233 (2022). <https://doi.org/10.1049/ntw2.12053>
- [28] Sambana, Bosubabu. (2020). An Artificial Intelligence based Air Pollution Prediction and Monitoring System using Internet of Things. *Xi'an Dianzi Keji Daxue Xuebao/Journal of Xidian University*. 14. 1634-1644. 10.37896/jxu14.7/187.
- [29] Goh, C.C.; Kamarudin, L.M.; Zakaria, A.; Nishizaki, H.; Ramli, N.; Mao, X.; Syed Zakaria, S.M.M.; Kanagaraj, E.; Abdull Sukor, A.S.; Elham, M.F. Real-Time In-Vehicle Air Quality Monitoring System Using Machine Learning Prediction Algorithm. *Sensors* 2021, 21, 4956. <https://doi.org/10.3390/s21154956>
- [30] Sofia, D.; Lotrecchiano, N.; Truccillo, P.; Giuliano, A.; Terrone, L. Novel Air Pollution Measurement System Based on Ethereum Blockchain. *J. Sens. Actuator Netw.* 2020, 9, 49.

# Optimized Strategy for Inter-Service Communication in Microservices

Sidath Weerasinghe, Indika Perera

Department of Computer Science & Engineering, University of Moratuwa, Sri Lanka

**Abstract**—In the last decade, many enterprises have moved their software deployments to the cloud. As a result of this transmission, the cloud providers stepped ahead and introduced various new technologies for their offerings. People cannot gain the expected advantages from cloud-based solutions merely by transferring monolithic architecture-based software to the cloud since the cloud is natively designed for lightweight artifacts. Nowadays, the end user requirements rapidly change. Hence, the software should accommodate those accordingly. On the contrary, with Monolithic architecture, meeting that requirement change based on extensibility, scalability, and modern software quality attributes is quite challenging. The software industry introduced microservice architecture to overcome such challenges. Therefore, most backend systems are designed using this architectural pattern. Microservices are designed as small services, and those services are deployed in the distributed environment. The main drawback of this architecture is introducing additional latency when communicating with the inter-services in the distributed environment. In this research, we have developed a solution to reduce the interservice communication latency and enhance the overall application performance in terms of throughput and response time. The developed solution uses an asynchronous communication pattern using the Redis Stream data structure to enable pub-sub communication between the services. This solution proved that the most straightforward implementation could enhance the overall application performance.

**Keywords**—Microservices; software architecture; inter-service communication; performance; streams

## I. INTRODUCTION

Recently, cloud computing has become considerably popular in the software industry, ultimately making businesses consider migrating their workloads to the cloud environment from their on-premise servers, as managing on-premise server farms is more costly and requires extra effort and maintenance. Based on the particular requirement, the consumers can choose the cloud services such as Infrastructure as a Service (IaaS), Platform as a Service (PaaS), software as a Service (SaaS), and Function as a service (FaaS). The main advantages of using the cloud are that consumers can gain whatever service they require according to their budget. The cloud provider fully manages the environment, and consumers no longer have the hassle of worrying about maintenance. Higher availability and easy vertical and horizontal scalability are some of the advantages of using cloud resources.

In the early days, requirements were very bounded, less volatile, and limited. Therefore, maintaining monolithic architecture software was easy. However, in modern society,

user requirements are complex and subject to constant change, making it cumbersome to make adequate changes to monolithic systems. In the monolithic architecture, all the data access layers, data store layer, logic layer, and user interface layer are tightly coupled into one single package. As a result, changing the code, adapting to the new technology, and testing the product have become problematic. Therefore, people invented Service Oriented Architecture (SOA) to design loosely coupled services. In modern SOA implementations, all the services are orchestrated by the Enterprise Service Bus (ESB). The main disadvantage of this architecture is that all the service calls are routed through this ESB, which is a single point of failure. The performance also tends to get impacted because of that [1].

Cloud services are designed for the light weighted small-scale artifacts like microservices. Hence, people cannot get the complete advantage of migrating their monolithic system into the cloud. Due to this cause, people are trying to reengineer their existing products into microservices architecture by separating the services and making them individual microservices [2]. There are many strategies to decompose the monolithic service into services, such as decomposing by the domains and subdomains, decomposing by the business capabilities, decomposing by the system responsibilities, and decomposing by the resources [3]. After converting monolithic applications into microservices architecture, people can gain many advantages such as improving code maintainability, adapting to new technologies, efficiently scaling only the required service, improving resilience, gaining more business agility, being easy to understand, etc. [4].

Some systems still have not moved their software to the microservice architecture because of certain performance issues related to response time and the application's throughput. Microservice design is an independent service; such services need to communicate with each other to provide the user requirements. Those services deployed in the distributed environment and for the communication services must send and receive the data packets through the network, adding extra latency compared to the monolithic architecture software. Synchronous and Asynchronous communication styles are the two communication styles that are used for microservices interservice communication [5]. Synchronous type still mainly uses the request/response-based behavior, and the request waits until the response reaches. Most people use HTTP or gRPC communication protocols due to inter-service communication in the microservices. Asynchronous communication styles use message brokers to exchange messages to the relevant microservices. They are mainly using the Pub/Sub mechanism,

which means that requests are not waiting for a response, and there are blocking threads associated with the communication. Researchers invert the broker less asynchronous methods, but there is no guarantee of the exact message delivery [6]. Most programming languages support microservice development, and they also develop the framework in conformance to that. Java language based Spring boot [7] and VertX [8], Node.js programming language based Molecular [9], and Golang programming language based GoMicro are fine examples of such instances [10].

This research focuses on the main performance issue of the microservice architecture, which is caused by the inter-service communication in the microservice architecture. As a result of the research, the solution was proposed to reduce the communication latency when communicating on microservice. The researcher has brought REST-based behavior to the top of the Redis Streams data structure for distributed communications. The network layer used the TCP-based socket connection and the serialized data packets to reduce the network latency while transferring the data. The rest of the paper discusses the implementation and the evaluation of this research outcome.

## II. LITERATURE REVIEW

### A. *Microservices*

Previous research publications showed that microservices research contributions started around the 2000 decade. After 15 years, Microservice research is drastically getting published in various academic journals and conferences [11]. Before the emergence of Microservice architecture, most engineers used Service Oriented Architecture (SOA) to build enterprise software. But, researchers have proven that they are faced with capacity issues and scaling issues with the SOA applications [12]. AI-Debagy and Martinek conducted a comparative review regarding the monolithic and microservice architecture. The experimental results of that research showed that monolithic applications perform 6% more on the throughput when compared to microservice-based applications [13]. Nevertheless, researchers have failed to elaborate on the underlying reason for the performance issue. The National Polytechnic School researched the challenges and problems faced during the system migration from monolithic to microservices architecture [14]. Finding suitable tools for migration, reorganizing the engineering team to work with the microservices, identifying the correct microservice design, guaranteeing consistency, and learning about the new framework are the challenges/problem they have highlighted in their research article. Lithuanian researchers reviewed the monolithic to microservice architecture, microservices methods and techniques [4]. One of the methods is to identify all subsystems associated with the monolithic architecture and create a dependency graph. Then architects can determine the services that need to be created as a microservice from the monolithic system. Another method is to identify the business logic from the dataflow diagram and decide what microservices can be created based on the independent business logic. As a best practice of the migration process, it is better to identify the minor steps and execute them one by one. With that, they have a guarantee on the path of restoration. The

software engineering department of Tashkent University published the mechanism to decompose the monolithic architecture system to microservice-based architecture with less development effort [15]. The process started with analyzing the monolithic system, then extracting micro functions, refactoring the service catalog, and finally, orchestrating the services. However, the researchers have not shown the proposed mechanism's real-world application. Florian Auer and team conducted the assignment to find out the facts that companies consider when migrating their system to microservice architecture [16]. Scalability, maintainability, complexity, reusability, modularity, deploy-ability, reliability and testability quality attributes of the Microservice are considered in their study. They have also figured out that most of the companies do not measure the process, product and the quality attribute in depth before the migration. Only after the migration that the companies realize the implications it has. This is caused because there is no standard framework and tech stack that engineer can use when developing the microservices.

### B. *Inter-Service Communication*

Presently, a lot of programming languages and microservice frameworks have emerged to develop microservices. When developing the Microservice, engineers use a tech stack solely based on their area of expertise. In most scenarios, they do not consider the application's nature. There are ample ways to perform service-to-service communication, which is the most crucial part of the microservice architecture. But, there is no clear-cut approach to identifying the most suitable and efficient method. Christy Pachikkal researched the microservices communication styles as synchronous communication and asynchronous messaging [5]. According to that research, developers need to intensely go through the system's functional / non-functional requirements and choose the correct communication style. Most developers use the REST protocol because of the ubiquity of the protocol, which makes them architecturally understand how this protocol works [17]. REST protocol mainly uses JSON format. But in certain instances, it uses the XML-based format for message passing. Those message formats take massive amounts of time to message parsing because of the weight of the message. As a solution, REST is supported for the binary JSON (BSON), which also has overhead with the field names within the data structure [18]. Google invented the Remote Procedure Call(RPC) framework-based protocol to get more performance than the REST over HTTP [19]. Abram Perdanaputra conducted research related to the microservice, which is deployed on the Kubernetes environment. In addition to that, all the communication is done by the gRPC protocol. They have decoded the request and the responses for transparent tracing, but that can be achieved in the passive mode. HTTP/2 was introduced in 2015. It is considered a binary protocol, which gives more efficient bandwidth usage and header compression [20]. Researchers have enabled the multiplexing for HTTP/2 protocol so that the clients can send multiple requests via the same TCP connection before the response is received by the client. This implies that if people can use the same TCP connection to send and receive messages, then they can reduce the latency in message passing. Google has invented a new protocol named Quick UDP Internet Connection (QUIC), which uses the User Datagram Protocol

(UDP) instead of the TCP connections, and behaves as a transport protocol for HTTP/2 [21]. Norwegian University researchers argue in other works that QUIC protocol performance degrades when the messages payload size gets larger than the HTTP/1.1 [22]. Gaetano Carlucci et al. conducted research on the QUIC and showed how QUIC and TCP protocols behave when the network is congested. Their experimental results have proven that TCP is able to provide better response time when compared to the QUIC protocol when a network packet loses.

A group of researchers in Indonesia has implemented asynchronous communication for the microservice architecture with the help of RabbitMQ message broker [23]. Seven Microservices were developed and deployed in an environment that could easily scale down. Communication between the services is done in an asynchronous event-driven manner which led to speed up the application because there was no waiting as request/response architecture. They proposed durable topics which can send the events to the subscriber, i.e., microservices when available. With this concept, they guarantee the message delivery to the client. Sanjana et al. researched the highly resilient inter-process communication service for the microservice architecture [24]. In their implementation, they have used the Kafka message broker and have enabled the pub/sub messaging style to do the inter-service communication. Researchers have used the Camel routes, which are capable of message transformations and validation when doing message routing. With that function, they have proven that inter-service communication can be done without changing the existing architecture of the microservice implementations. Therefore, they argue that provided solution is highly resilient and lightweight for inter-service communication. This non-functional requirement is brought up because they have implemented the solution over the existing framework.

### III. METHODOLOGY

This research focuses on implementing a solution for the inter-service communication method to improve the overall performance in a microservice architecture. When implementing the solution, the researcher has considered asynchronous communication as a communication style according to the facts found in the literature review part. The proposed system uses the Redis with Streams data structure, combined with Pub/Sub communication pattern for message passing in the underlying implementation.

#### A. Redis

Redis is an open-source solution that people can use as an in-memory data structure store. The advantage of the Redis is that it gives high read/write speed with high concurrency [25]. Redis is, by default, a support for easy scaling with the cluster concept. It also brings high availability and fault tolerance with the concept of virtual hash slots. Redis uses its own communication protocol to engage with clients as RESP (Redis Serialization Protocol), and it is also a binary safe protocol [26]. Since it is a Serialization communication protocol with binary safety, transporting the payloads will take less bandwidth. All the clients are connected to the Redis using the

TCP connection. In this proposed method, the stream-oriented connection, which is similar to Unix sockets is used.

#### B. Redis Streams

Redis streams are introduced from Redis 5.0, which can publish the message to the stream, and consumers who subscribe to that stream can receive them. Redis streams differ from Kafka because the stream is an append-only data structure that helps with real-time messaging. The main advantage over the pub/sub model is that Redis streams persist the messages. Hence, it can guarantee the exact message delivery. Message reliability is the most important part of microservice inter-service communication, and can be achieved from this model.

#### C. Component Architecture

Fig. 1 depicts the system architecture of the proposed solution. Microservice A, B, and C are independent microservices that are deployed in the distributed environment. In order to produce the functional requirements in the software, each microservice needs to communicate with one another. In the proposed solution, communication is enabled through the Redis Streams as Pub/Sub communication style. Every microservices creates a Unix-based socket connection from the microservice to the Redis server via TCP 6379 port. All the messages are passed through that TCP connection over the RESP protocol. Every time the microservice does not create and close the connection, when the microservice starts, the TCP connection creates and will live until it shuts down. Thus, network creation and closing time can be reduced with this approach. Using this mechanism developed, the Java-based library enables efficient communication between the microservices and brings all attributes of the HTTP protocol to the developed library. Programmers can use this without changing the existing architecture of their system.

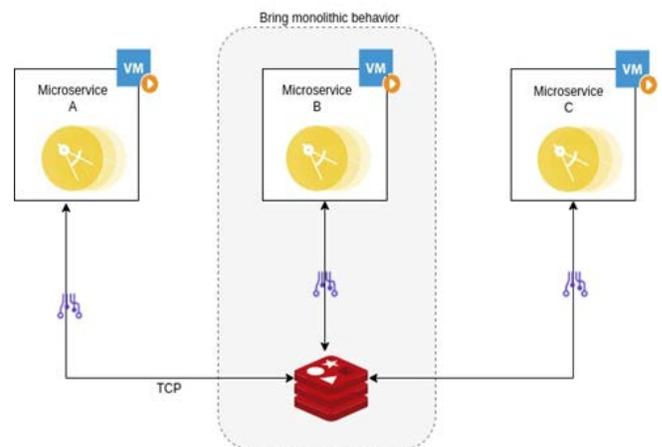


Fig. 1. High-level system architecture

### IV. IMPLEMENTATION

This section briefly describes the implementation of the proposed solution. Java programming language and Spring boot microservice framework have been used to implement the proposed solution according to the literature review the researcher has conducted. The researcher has implemented the request/response-based stateless client like HTTP Client, but beneath, it works on the pub/sub communication style. As a

communication medium, the researcher has used the Redis Streams. Spring Boot provides a spring-boot-starter-data-redis library, which can be used to build the solution [27]. That gives high-level and low-level abstractions for integrating with the Redis server.

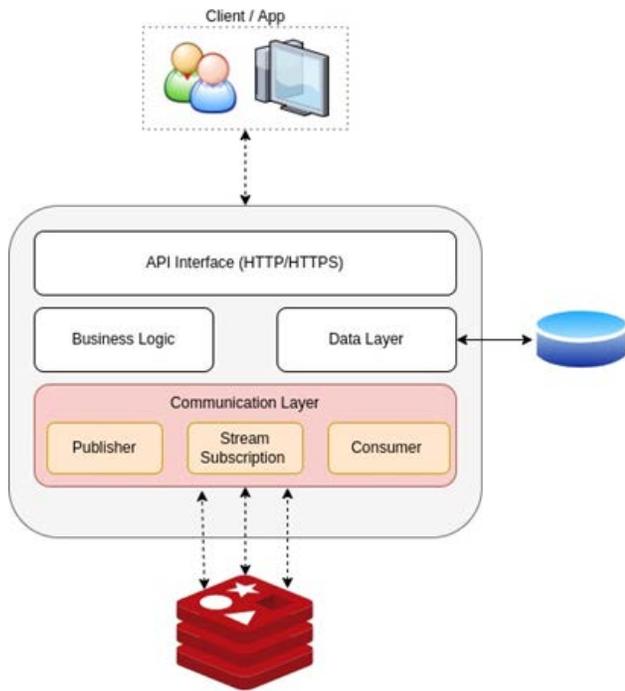


Fig. 2. High-level component architecture

The researcher has segregated one microservice into four main categories (see Fig. 2). The main category is the API interface, the programming component that the microservice communicates with external parties. In this implementation, the interface is neither touched nor changed because the external parties use that and cannot adhere to the implementations by changing their applications. Other categories are the business logic part and the data access layer. Business logic is the main part that contains the functionalities of the microservices. Most of the microservices are segregated from these business functions. Another category is the data layer, the part that communicates with the data sources, such as the database. We have developed the communication layer with three sub-programming components: publisher, stream subscription, and consumer.

- 1) *Publisher*: Responsible for sending the message to the correct microservice.
- 2) *Stream subscription*: Decides the destinations of the messages.
- 3) *Consumer*: Responsible for receiving messages.

#### A. Request Handling in Microservices

3<sup>rd</sup> party client sends the HTTP request by invoking the API exposed from the microservice A, as per Fig. 3. The researcher has used Spring Boot REST Controller to expose the API, which gives the developer the to enable restful web services. When the microservice starts, it establishes the Unix socket-based connection to the Redis server using the TCP port

with the configured IP and the port. After that, create the subscription using the stream keys, which are also configured in the properties file. Stream keys belong to the other microservices that microservice A needs to send the messages. After receiving the request from the 3<sup>rd</sup> party client, microservice A starts processing that request and makes the EventStructure object using the request details and processed details. EventStructure is the object sent as a message to microservice B to get more details. That object contains all the HTTP message details such as HTTP method, parameters, headers, body, client details, publisher details, etc. After processing the HTTP request, the microservice decides which microservice needs to be called to provide the client's correct response. Based on that, microservice A chooses the correct stream key and publishes the message. When publishing the EventStructure object, it is serialized and passed as a byte buffer record. Because of this mechanism, network consumption can be vastly reduced when transferring data. Each published event has a unique event ID. Afterward, that ID and EventStructure object will be stored as a callback reference in a HashMap to process the response.

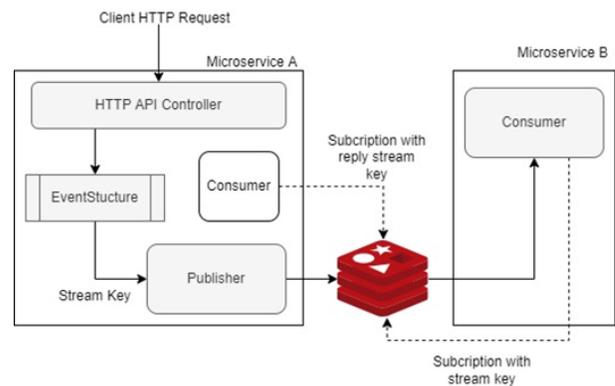


Fig. 3. How request serve

#### B. Response Handling in Microservices

After processing, the business logic response can be set to the EventStructure object. Publisher and reply stream key data can be retrieved from the EventStructure object and can publish the response using those data. Consumers placed in the microservice A can identify the client data by receiving the stream event and mapping the response data with the HashMap data, which is stored earlier. After processing, the response API controller can send the response back to the client using the HTTP protocol as per Fig. 4.

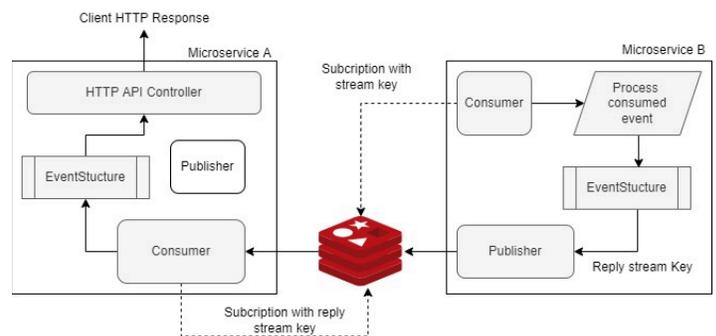


Fig. 4. How response serve

### C. Quality Attributes

Quality attributes are the most vital in Microservice architecture. Most people transition to microservice architecture considering scalability, performance, maintainability, traceability, and availability [28]. When implementing the new solution, the quality attributes are preserved and improved.

1) *Scalability*: Proposed solution can be horizontally (scaling out) and vertically (scaling up) scalable. If this solution is deployed in the cloud-based VM, the VM specification can be increased at any given time and ultimately support vertical scaling. Thus, JVM has more resources to execute computations, and there is no barrier to the implantation. By performing horizontal scaling, people can add more microservice instances based on business needs. When adding the new application, it creates the subscription using the stream key. Redis server is responsible for delivering the messages solely to one subscriber, which means that the request is received only by one microservice. The Redis service covers service discovery and load balancing. Therefore, the developer does not need to ponder on it when scaling the applications. Hence, message duplication is not happening with this implementation, guaranteeing the exact message's delivery.

2) *Maintainability*: There is no impact on the overall maintainability of this implementation. Developers can use this implementation for internal communication instead of HTTP Clients. There's no requirement for maintenance in the internal load balancer for inter-service communication.

3) *Traceability*: This is the most important quality attribute in relation to technical support, as the troubleshooting support engineers need to know what has happened to the request and the response. The developed implementation supports end-to-end request/response tracing via the Redis server. If there's a need to trace the request and the response, the Redis GUI client can be installed after connecting to the Redis server of that client.

4) *Availability*: With this implementation, exact message delivery is guaranteed from the Redis Streams. Hence, availability can be achieved through this.

## V. RESULTS AND DISCUSSION

By critically reviewing the microservice architecture, the researcher analyzed that the impact of inter-service communication on performance is very high as a result of all the microservices deployed in the distrusted environment, making it mandatory to call each other over the network to produce the results. In this research, the researcher has proposed and implemented a solution that can be used to improve inter-service communication, ultimately bringing in overall application performance in terms of response time and throughput. HTTP inter-service communication and implemented solution have been deployed in a cloud VM-based environment to test and evaluate the application response time and throughput.

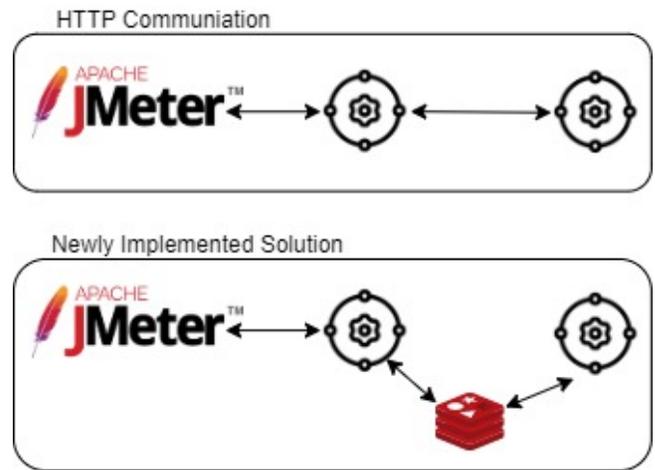


Fig. 5. Testing architecture

The Fig. 5 depicts the high-level system architecture of the load-testing environment. To generate the load, the researcher has used Apache JMeter, which is the most famous and vastly used tool in the industry, as well as in academic research [29]. Well-known cloud provider AWS (Amazon Web Services) has been used to deploy the systems [30]. Most enterprise software companies and enterprise-grade software are developed in the AWS cloud. Numerous scientific types of research are also conducted recently from the AWS cloud. Hence, the AWS platform is chosen in this instance as well to evaluate the system [31]. AWS EC2 is a virtual machine infrastructure as a service that contains the Intel Xeon processors with burstable for high frequency and a balanced memory/network and IO resources.

The T2 instance type has been chosen as it is a low-cost general-purpose instance category that provides better CPU performance for microservices and low-latency interactive applications [32]. T2.Medium EC2 instance type has been used to deploy the two microservices and JMeter, which has two virtual CPUs and 4GB memory on each. T2. A small instance type is used to deploy the Redis server, which contains one virtual CPU and 2GB of memory. All the VMs are provisioned in one virtual private network (VPN) and the same subnet under one security group. This network architecture can minimize network latency by calling the application through the same subnet and improves security by using the same security group.

In Fig. 6, straight arrows are the path that conducts the test for the common standard HTTP communication. The dotted arrows depict the newly implemented solution load test path.

The test is executed in two different methodologies,

1) *Scenario A*: Controlled the application's overall throughput and the request/response size and then measured the inter-service communication turnaround time.

2) *Scenario B*: Controlled only the request/response size and measured the throughput, overall application response time, and inter-service communication turnaround time.

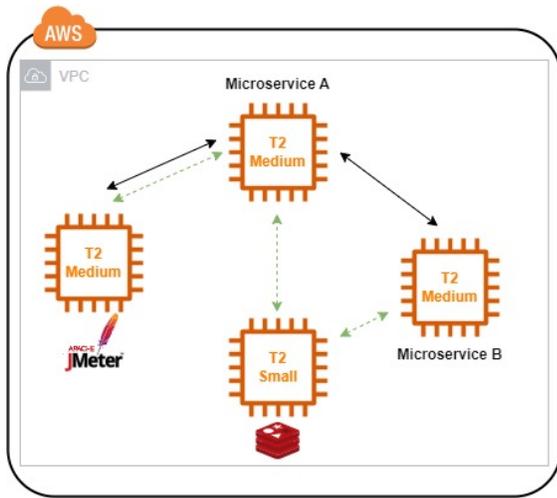


Fig. 6. Deployment architecture

A. Scenario A

Based on the studies conducted, the researcher has chosen different testing scopes to evaluate the system with existing systems. The researcher started with different throughput values and payload sizes. For each constant throughput value, the payload size has been changed as below, and traffic has been generated from the JMeter.

- Call the HTTP GET method from the microservice, and the backend microservice returns the 200 OK HTTP response code with the empty body.

Call the HTTP POST method with a 1KB size JSON payload, and the backend microservice returns the 200 OK HTTP response with a 1KB JSON format response.

Using the above test scenarios (Table I), the researcher has evaluated the HTTP communication method, inter-service communication turnaround time, and the newly implemented solution for inter-service communication turnaround time. Each test scenario was run for a time period of 1 hour and repeated three times, generating an average value of the inter-service communication.

Fig. 7 reflects the difference between the turnaround time on the proposed solution and the HTTP protocol implementations. When the throughput gets high, both the proposed solution and the HTTP protocol solution's turnaround time increase; when considering the payload size, it can be comprehended that both perform the same behavior. But in the all-test scenarios, the proposed solution's turnaround time is getting much lower than HTTP protocol implementation. In the HTTP protocol, a socket connection needs to be created for each connection to close the connection once the response is received. Furthermore, the network packets are neither in a binary method nor fully serialized. Therefore, when transferring the data packet through the network consumes considerable time. In the new implementation, Microservices has established a TCP socket-based connection with the Redis server. Hence, when Microservices starts, it acts as part of the particular Microservice. When sending data to other Microservices, it needs to be serialized and passed as a byte buffer record to reduce the usage of network resources. Due to

that, the implemented solution turnaround time is less than the standard HTTP communication method.

B. Scenario B

In this scenario, only the payload sizes have been controlled and evaluated for the application's overall response time and the Microservice's inter-service communication turnaround time. The researcher has conducted this test scenario in the same cloud environment, and for the payload size, only 1KB sized JSON payload, 5KB sized JSON payload, and URL were chosen. To capture the overall application response time, JMeter listeners are being added. The inter-service communication turnaround time has been calculated by processing the logs. Each test scenario was run for 1h and continued three times to get the average value of the inter-service communication turnaround time, overall application response time, and application throughput.

TABLE I. TEST SCENARIOS

Number	Test Case Scenario
1	Controlled the throughput to 10TPS and send HTTP GET request
2	Controlled the throughput to 10TPS and send HTTP POST request
3	Controlled the throughput to 100TPS and send HTTP GET request
4	Controlled the throughput to 100TPS and send HTTP POST request

Turnaround Time Comparison

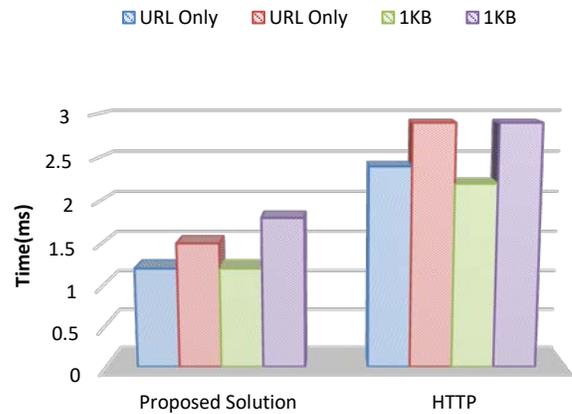


Fig. 7. Turnaround time comparison chart

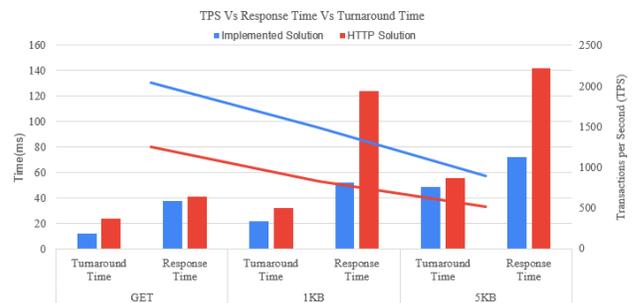


Fig. 8. TPS / response time / turnaround time comparison chart

As in Fig. 8 it can be observed that with the increase of the payload size, throughput is getting decreased. It is a typical network behavior that, when packets get heavy, will decrease the overall network performance. Hence, the response time also increases when the payload size increases. By comparing the implemented solution and the generic HTTP interservice communication method, it can be seen that in all the cases, response time gets better in implemented solution compared to the HTTP communication method.

Critically evaluating the above diagram, it can be concluded that;

Inter – service communication turnaround time  
 $\propto$  Application response time

Inter-service communication turnaround directly impacting to the whole application response time. This means that if some systems took more time to communicate between services, then overall response time will become high on that system due to inter-service communication.

Payload size  $\propto$  1/Throughput

Request and response payload size impacting to the system throughput because transferring large network packets will take some considerable time between services. Hence response time will be getting increased. With the results of that, overall system throughput will be getting decreased.

## VI. CONCLUSION AND FUTURE WORK

In Microservice architecture, all the services are deployed as independent services in a distributed environment. However, unlike in monolithic software, the data must be derived through the network call to share the data between services. As a result of that, additional latency will be added to the overall application response time. Most software companies are faced with issues related to performance in terms of response time and throughput when migrating their monolithic architecture to microservice-based architecture. However, there is a capacity-wise and cost-wise advantage by scaling required services when necessary.

This research focuses on finding a solution to reduce the inter-service communication time between services. The initial studies found that most of the existing protocols take time for connection establishment and connection closure when sending and receiving the response. Besides, sending massive payloads will cause additional latencies. We have implemented the solution by addressing the above-mentioned problems and reducing latency when communicating between the services. We have used the Redis Stream data structure and built the request/response-based message-passing solution for inter-service communication. A TCP-based socket connection is created when the microservice starts. When sending the payload, it will be serialized and sent as a protocol buffer. Redis server is responsible for the exact message delivery based on the subscription and the stream key. The test scenarios are conducted by deploying the implemented solution in the AWS cloud-based VMs, and the system is evaluated against the Spring Boot standard implementation. Test results depict that the implemented solution performs well in terms of application response time and throughput. This research will

continue to find a cloud-native solution to gain more performance and maintainability.

## REFERENCES

- [1] S. Weerasinghe and I. Perera, "An exploratory evaluation of replacing ESB with microservices in service oriented architecture," presented at the International Research Conference on Smart Computing and Systems Engineering, Sep. 2021.
- [2] A. Makris, K. Tserpes, and T. Varvarigou, "Transition from monolithic to microservice-based applications. Challenges from the developer perspective," *Open Res. Eur.*, vol. 2, p. 24, Feb. 2022, doi: 10.12688/openreseurope.14505.1.
- [3] Chris Richardson, *Microservices patterns*. Manning Publications, 2018.
- [4] J. Kazanavicius and D. Mazeika, "Migrating Legacy Software to Microservices Architecture," in 2019 Open Conference of Electrical, Electronic and Information Sciences (eStream), Vilnius, Lithuania, Apr. 2019, pp. 1–5. doi: 10.1109/eStream.2019.8732170.
- [5] Christy Sibi Pachikkal, "Interservice Communication in Microservices," *Int. J. Adv. Res. Sci. Commun. Technol.*
- [6] S. Raje, "Performance Comparison of Message Queue Methods", doi: 10.34917/16076287.
- [7] "Spring Boot." <https://spring.io/projects/spring-boot>.
- [8] "Eclipse Vert.x." <https://vertx.io/>.
- [9] "Moleculer - Progressive microservices framework for Node.js," Moleculer - Progressive microservices framework for Node.js. <https://moleculer.services/index.html>.
- [10] A. Aslam, "Go Micro." [Online]. Available: <https://github.com/asim/gomicro>.
- [11] Sidath Weerasinghe and Indika Perera, "Taxonomical Classification and Systematic Review on Microservices," *Int. J. Eng. Trends Technol. - IJETT*, Accessed: Jul. 30, 2022. [Online]. Available: <https://ijettjournal.org/archive/ijett-v70i3p225>.
- [12] L. D. S. B. Weerasinghe and I. Perera, "An exploratory evaluation of replacing ESB with microservices in service-oriented architecture," in 2021 International Research Conference on Smart Computing and Systems Engineering (SCSE), Sep. 2021, vol. 4, pp. 137–144. doi: 10.1109/SCSE53661.2021.9568289.
- [13] O. Al-Debagy and P. Martinek, "A Comparative Review of Microservices and Monolithic Architectures," in 2018 IEEE 18th International Symposium on Computational Intelligence and Informatics (CINTI), Budapest, Hungary, Nov. 2018, pp. 000149–000154. doi: 10.1109/CINTI.2018.8928192.
- [14] V. Velepucha and P. Flores, "Monoliths to microservices - Migration Problems and Challenges: A SMS," in 2021 Second International Conference on Information Systems and Software Technologies (ICI2ST), Quito, Ecuador, Mar. 2021, pp. 135–142. doi: 10.1109/ICI2ST51859.2021.00027.
- [15] D. Kuryazov, D. Jabborov, and B. Khujamuratov, "Towards Decomposing Monolithic Applications into Microservices," in 2020 IEEE 14th International Conference on Application of Information and Communication Technologies (AICT), Tashkent, Uzbekistan, Oct. 2020, pp. 1–4. doi: 10.1109/AICT50176.2020.9368571.
- [16] F. Auer, V. Lenarduzzi, M. Felderer, and D. Taibi, "From monolithic systems to Microservices: An assessment framework," *Inf. Softw. Technol.*, vol. 137, p. 106600, Sep. 2021, doi: 10.1016/j.infsof.2021.106600.
- [17] O. Zimmermann, "Microservices tenets: Agile approach to service development and deployment," *Comput. Sci. - Res. Dev.*, vol. 32, no. 3–4, pp. 301–310, Jul. 2017, doi: 10.1007/s00450-016-0337-0.
- [18] M. Ya. Afanasev, Y. V. Fedosov, A. A. Krylova, and S. A. Shorokhov, "Performance evaluation of the message queue protocols to transfer binary JSON in a distributed CNC system," in 2017 IEEE 15th International Conference on Industrial Informatics (INDIN), Jul. 2017, pp. 357–362. doi: 10.1109/INDIN.2017.8104798.
- [19] L. N. T. Thanh, "SIP-MBA: A Secure IoT Platform with Brokerless and Micro-service Architecture," *Int. J. Adv. Comput. Sci. Appl.*, vol. 12, no. 7, p. 8, 2021.

- [20] R. Corbel, E. Stephan, and N. Omnes, "HTTP/1.1 pipelining vs HTTP2 in-the-clear: Performance comparison," in 2016 13th International Conference on New Technologies for Distributed Systems (NOTERE), Jul. 2016, pp. 1–6. doi: 10.1109/NOTERE.2016.7745823.
- [21] H. Bakri, C. Allison, A. Miller, and I. Oliver, "HTTP/2 and QUIC for Virtual Worlds and the 3D Web?," *Procedia Comput. Sci.*, vol. 56, pp. 242–251, Jan. 2015, doi: 10.1016/j.procs.2015.07.204.
- [22] M. S. Nyfløtt, "Optimizing Inter-Service Communication Between Microservices," p. 103.
- [23] S. A. Asri, I. N. G. A. Astawa, I. G. A. M. Sunaya, I. M. R. Adi Nugroho, and W. Setiawan, "Implementation of Asynchronous Microservices Architecture on Smart Village Application," *Int. J. Adv. Sci. Eng. Inf. Technol.*, vol. 12, no. 3, p. 1236, Jun. 2022, doi: 10.18517/ijaseit.12.3.13897.
- [24] S. G. B. and G. R. S. N. S., "High Resilient Messaging Service for Microservice Architecture," *Int. J. Appl. Eng. Res.*, vol. 16, no. 5, p. 357, May 2021, doi: 10.37622/IJAER/16.5.2021.357-361.
- [25] X. Chen, F. Wang, J. Xu, D. Zhu, P. Tan, and J. Ma, "A distributed cache system based on Redis for high-speed railway catenary monitoring system," in 2020 Chinese Automation Congress (CAC), Shanghai, China, Nov. 2020, pp. 2048–2053. doi: 10.1109/CAC51589.2020.9326531.
- [26] "RESP protocol spec," Redis. <https://redis.io/docs/reference/protocol-spec/>.
- [27] "Spring Data Redis." <https://spring.io/projects/spring-data-redis>.
- [28] T. Schirgi, "Architectural Quality Attributes for the Microservices of CaRE," p. 46.
- [29] R. B. Khan, "Comparative Study of Performance Testing Tools: Apache JMeter and HP LoadRunner," p. 57.
- [30] "AWS Lambda – Serverless Compute - Amazon Web Services," Amazon Web Services, Inc. <https://aws.amazon.com/lambda/>.
- [31] "Security and Safety in Amazon EC2 Service – A Research on EC2 Service AMIs," *Int. J. Innov. Technol. Explor. Eng.*, vol. 8, no. 6S4, pp. 736–738, Jul. 2019, doi: 10.35940/ijitee.F1149.0486S419.
- [32] "Amazon EC2 T2 Instances – Amazon Web Services (AWS)." <https://aws.amazon.com/ec2/instance-types/t2/>.

# Deep Learning based Analysis of MRI Images for Brain Tumor Diagnosis

Srinivasarao Gajula<sup>1</sup>, V. Rajesh<sup>2</sup>

Research Scholar<sup>1</sup>, Professor<sup>2</sup>

Department of ECE, Koneru Lakshmaiah Education Foundation  
Guntur, AP, India-522302<sup>1,2</sup>

**Abstract**—This Identification and examination of brain tumour are critical components of any indication system, as evidenced by extensive research and methodological advancement over the years. As part of this approach, an efficient automated system must be put in place to enhance the rate of tumor identification. Today, manually examining thousands of MRI images to locate a brain tumor is arduous and imprecise. It may impair patient care. Since it incorporates several picture datasets, it might be time-consuming. Tumor cells present in the brain look a lot like healthy tissue, making it hard to distinguish between the two while doing segmentation. In this study, we present an approach for classification and prediction of MRI images of the brain using a convolutional neural network, conventional classifiers, and deep learning. Here we have proposed a new method for the automatic and exact categorization of brain tumour utilizing a two-stage feature composition of deep convolutional neural networks (CNNs). We used a deep learning approach to categorize MRI scans into several pathologies, including gliomas, meningiomas, benign lesions, and pituitary tumour, after first extracting characteristics from the scans. Additionally, the most accurate classifier is selected from a pool of five possible classifiers. The principal components analysis (PCA) is used to identify the most important characteristics from the retrieved features, which are then used to train the classifier. We develop our proposed model in Python, utilizing TensorFlow and Keras since it is an effective language for programming and performing work quickly. In our work, CNN got a 98.6% accuracy rate, which is better than what has been done so far.

**Keywords**—Convolutional neural networks (CNN); magnetic resonance imaging (MRI); principal components analysis (PCA)

## I. INTRODUCTION

In the field of medicine, images need to be segmented to be broken down into their constituent pieces. To better analyze a picture, it is helpful to "spit out" the representation of the image. This is because the picture is segmented into several individual parts. Recently, deep learning models have begun to make their presence known in the world of biomedical applications. The network that is used for deep learning has many layers that are hidden from view.

Brain tumour is one of the most debilitating diseases. Brain tumours are frequently found in several locations in the brain. The diagnosis of brain tumour at an earlier stage is essential for more effective treatment. After a clinical suspicion of a brain tumour, imaging studies are necessary to pinpoint the exact location of the tumour, measure its size, and assess its potential

for spreading to other parts of the brain and skull. Based on these findings, doctors select the most effective course of treatment from among surgery, radiation, and chemotherapy. The survival rate of a patient afflicted with a tumour can obviously be raised with precise early detection. Therefore, the radiology department has placed a greater emphasis on the research of brain malignancies through imaging modalities.

Brain tumours are malignant growths that start in the brain or the protective membranes that surround the brain and skull. Finding malignant tumours anywhere on the body and treating them early is challenging. According to a current study, the incidence of brain tumours has grown significantly. Problems with hearing or speech, persistent headaches, memory decline, vision loss, and behavioral shifts are all indicators of a more serious brain disorder. The performance of image processing steps depends heavily on the results of image segmentation [1]. In this situation, we have been particularly concerned with extracting the tumour from the MRI scans of the brain. Medical professionals can pinpoint the exact site of the brain tumour with greater precision. Radiologists, engineers, and physicians all use medical image processing to gain a deeper understanding of patients. Considering the difficulties in segmenting brain tumours and the importance of this task in clinical practice, a wide range of segmentation, automation, and semi-automation mechanisms have been developed. Deep learning techniques rule medical image processing techniques and research [2]. The CNN is a classification method that uses deep learning to determine different types of brain tumors.

The different analyses were performed using deep learning techniques to segment and detect brain tumors. For brain tumour segmentation, Yanming Sun et al. [3] proposed a robust and efficient CNN technique. John Schmeelk [4] worked with two-dimensional images by employing a two-dimensional wavelet transform. To complete the procedure of segmenting MRI brain images, Parra et al. [5] utilised an artificial neural network (ANN) method. In this study, a learning vector quantization (LVQ) network was constructed using an ANN method. Papageorgiou et al. [6] used fuzzy cognitive maps (FCMs) to describe design specialists. Incorporating a computationally clever training strategy known as the activation Hebbian algorithm enhanced the FCM ranking model's classification accuracy. El-Sayed et al. recently introduced decision support systems in medicine that use normal and abnormal categories. MRI scans were classified using a hybrid framework [7]. There are three main phases to the hybrid architecture that is proposed. In the first step, MRI

scans are extracted using discrete wavelet transformations (DWT). In the second step to minimize the number of MR image features, we used principal component analysis (PCA). In the end, two different classifiers were utilized to differentiate normal MR images from those that were aberrant. Othman et al. used a probabilistic neural network (PNN) and image processing to classify brain tumours automatically [8]. The suggested PNN classifier went through the decision-making process in two stages. The first stage involved the extraction of features via the use of principle component analysis (PCA), and the second stage involved classification through PNN. Pei et al. suggested a method that improves texture-based tumour segmentation in longitudinal MRI by utilizing tumour growth patterns as novel characteristics. This method makes use of tumour growth patterns [9]. A new convolutional neural network (CNN) model for brain tumour classification was proposed by Muhammad Sajjad et al [10].

The MR images were segmented to determine where the tumour was located. The next step was to enrich the existing data set. Next, they used the proposed CNN to carry out the categorization procedure. They have a 94.58% success rate when classifying data. The MR dataset was split into two categories, normal and aberrant, by Kanmani et al. [11]. To improve classification efficiency, they applied the threshold-based region optimization (TBRO) technique. They used this technique to carry out segmentation. The proposed method was experimentally validated with a success percentage of 96.57 percent. Seetha J. et al. [13] proposed an approach to speed up computations while maintaining high precision. The ImageNet database is utilised for categorization purposes. To obtain high accuracy, a loss function that is based on gradient descent is implemented. Fuzzy C-Means clustering was used for tumour segmentation by Tonmoy Hossain et al. [14], and it accurately predicted tumour cells. After performing segmentation, they applied traditional classifiers and a convolutional neural network to choose features for classification. In the traditional classifier section, they applied and compared the outcomes of various traditional classifiers such as K-Nearest Neighbor, Logistic Regression, Multilayer Perceptron, Nave Bayes, Random Forest, and Support Vector Machine. A comprehensive overview of existing techniques for segmenting and identifying MRI brain images was provided by Srinivasarao Gajula et al. [15]. You may learn in this article how many people are diagnosed with brain tumours annually. General advice about how to stay healthy while dealing with the sickness is included [16]. The modality-paired learning approach, which employs a 3D U-Net as its backbone network, was proposed by Yixin Wang et al. [17]. This method employs paralleled branches to independently extract features from multiple modalities before combining them via efficient layer connections. Gajula, S., et al. [18] proposed that the machine-learning technique of logistic regression may be used to do automatic brain abnormality identification. Disorders including Alzheimer's, transient ischemic attacks (TIAs), and brain tumours are all identifiable by this TSLR model. The most frequently accessed data sources were reported by Wenyin Zhang et al. [19]. The authors then provide a brief overview of the three types of multi-modal brain tumour MRI image segmentation techniques: traditional segmentation techniques, segmentation techniques based on classical machine learning

techniques, and segmentation techniques based on deep learning techniques. Typical algorithmic structures, principles, and benefits and drawbacks are outlined for each approach. At last, they discuss the difficulties and offer a possible perspective for future patterns of development. Gupta, Gaurav, and colleagues [20] proposed data mining techniques for MRI image classification. This research proposes a novel hybrid approach to brain tumour classification using support vector machines (SVM) and fuzzy c-means. Image-enhancing methods including contrast enhancement and mid-range stretching are implemented in this algorithm. Skull stripping uses a combination of double-thresholding and morphological processes. To find the suspicious area in a brain MRI, fuzzy c-means (FCM) clustering is applied to the segmented image. Brain MRI images are classified using support vector machines (SVMs) after features are extracted using a grey-level run-length matrix (GLRLM). A modified KNN-based clustering and segmentation method is proposed by Xie et al. [21] that uses Minkowski distance as the primary parameter. Using transfer learning, Srinivasarao Gajula et al. [22] presented the super pixel method for detecting and segmenting brain tumours. Data sets were pre-processed initially before being fed into a VGG-19 transfer learning network, which was then used to detect brain tumours. The next step is for a UNet model to locate malignant growths. In this study, we offer an effective and skilled technique based on both conventional classifiers and convolutional neural networks for automatically segmenting and detecting brain tumours.

The focus of this research is on the use of magnetic resonance imaging (MRI) of the brain to detect tumors. Finding brain tumours early is important because it helps medical professionals make a correct clinical diagnosis. The purpose of this work is to develop an algorithm for detecting tumours in MR brain images that is both accurate and reliable. Neurosurgeons and other medical professionals can use the system. The technology, which makes use of image processing, pattern analysis, and computer vision, is designed to boost the sensitivity, specificity, and efficiency of screening for brain tumours.

This paper is organized as follows: Section II of this study presents the proposed methodology that will be discussed in detail. Section III presents the results and discussion, and Section IV presents the conclusion and future work.

## II. PROPOSED METHODOLOGY

Using a convolutional neural network, which we show here, we provide a method for the categorization and prediction of MRI images of the brain. Our initial prospective model classified and identified brain tumours using machine learning methods. The suggested approach aims to improve human health and longevity by classifying brain cancers. The suggested effort attempts to simplify the classification of brain tumours while simultaneously increasing their accuracy compared to existing methods. This technique was selected because, in comparison to conventional CNN, it possesses greater capabilities in terms of both accuracy and speed when it comes to classification.

### A. Data Acquisition

The dataset that was gathered is separated into two categories: training brain images and testing brain images. Among these training datasets contains 2870 images of 4 different classes like pituitary, meningioma, glioma, and no tumor, and the testing dataset contains 394 images of 4 different classes. These images are analyzed, and then the algorithms get processed once they've been pre-processed.

### B. Pre-Processing

During the pre-processing step, the primary objective is to precisely eliminate the redundancy that was present in the image that was collected while maintaining the details that are an important part of the entire method. This is done to improve not just the quality of an image but also its entire appearance. Removing unwanted noise from an image is a key part of any pre-processing method used to fix a degraded image [12]. Adaptive filtering is one kind of denoising in which the process is carried out in accordance with the noise data already present in an image on a regional level. Possible image de-noising methods are as follows:

$$I(x, y) = \frac{\sigma_y^2}{\sigma_{local}^2} (I(x, y) - \mu_{local}) \quad (1)$$

Where  $I(x, y)$  is the reduced image,

$\sigma_y^2$  =variance

$\mu_{local}$  = mean around window pixel

$\sigma_{local}^2$  = local variance of window.

### C. Segmentation

The process of segmentation is used with several medical imaging modalities to identify contaminated tumour tissue. Image analysis must begin with the stage of segmentation because it is essential to the process. Segmentation is the process of dividing an image into distinct sections or blocks that have similar and identical characteristics. The procedure of segmenting a brain tumour involves separating the normal brain tissues and solid tumours from the tumour tissues, such as hydrocephalus and dead cells, which are found within the tumour. Image segmentation has several methods. The choice of segmentation techniques, on the other hand, is determined by the kinds of features that are going to be processed and extracted.

### D. Global Threshold Segmentation:

A threshold image can be given as  $g(x, y)$

$$\text{Here } g(x, y) = \begin{cases} 1, & \text{if } f(x, y) > T \\ 0, & \text{if } f(x, y) < T \end{cases} \quad (2)$$

here 1 is object and 0 is background and  $f(x, y)$  is input image. In case of global thresholding T is constant applicable over whole image. Algorithm for automatic estimation of threshold T is as follows.

*Step 1:* Select an initial estimate for T.

*Step 2:* Segment image using T as two groups with group1 (G1) values as greater than T and group 2 (G2) values as less than or equal to T

*Step 3:* Compute average intensity values for G1 as m1 and for G2 as m2

*Step 4:* Compute a new threshold value  $T_{new} = 1/2 (m1+m2)$

*Step 5:* Repeat (2) through (4) until the difference in  $T_{new}$  in successive iterations is smaller than  $\Delta T$ .

In the past, the most prevalent segmentation methods were pre-processing and thresholding, or a mixture of the two. The thresholding method is the simplest, and it incurs the least amount of calculation cost. The histogram of the image is an essential factor in determining the global threshold. Utilizing the values of the local attributes allows for not only the enhancement of the histogram but also the computation of the global threshold to be performed. At the beginning of this procedure, we will select the MRI image of the brain. In the following stage, we will be calculating the value of the threshold. After that, we will have two distinct groupings of pixel values. We are computing the arithmetic mean by utilizing these two groups. The new threshold is determined by arithmetically averaging the two means. This technique is repeated until the desired number of iterations has been reached.

### E. Feature Extraction

The term "feature extraction" is used to describe the procedure of reducing unstructured data to a set of quantifiable characteristics that may then be processed without losing any of the original data's context. To better characterize a large dataset with fewer resources, feature extraction is employed. To classify the data, two different kinds of features were extracted: texture-based features and statistically based features. During the process of feature extraction, we will obtain several characteristics of the images, such as their mean, skewness, entropy, standard deviation, centroid, energy, dissimilarity, homogeneity, and correlation features.

### F. Proposed Methodology using CNN

To detect tumours, a five-layer convolutional neural network has been developed and is currently being used. In our 5-layer CNN model, there are seventeen stages, as well as the hidden layers, which provide us with the best possible outcome in terms of tumour detection. In the domain of medical image processing, convolutional neural networks have found widespread application. Throughout the years, several researchers have attempted to construct a model that can detect the tumour in a more effective manner. We attempted to develop a model that is capable of accurately classifying the tumour based on images taken from a 2D MRI of the brain. Although a fully connected neural network is capable of tumour detection, we elected to employ a convolutional neural network (CNN) for our model due to parameter sharing and sparsity of connection.

The process by which the suggested work will be carried out on the chosen data sets is depicted for us in Fig. 1. Fig. 2 displays a variety of MRI images, some of which contain tumours while others do not.

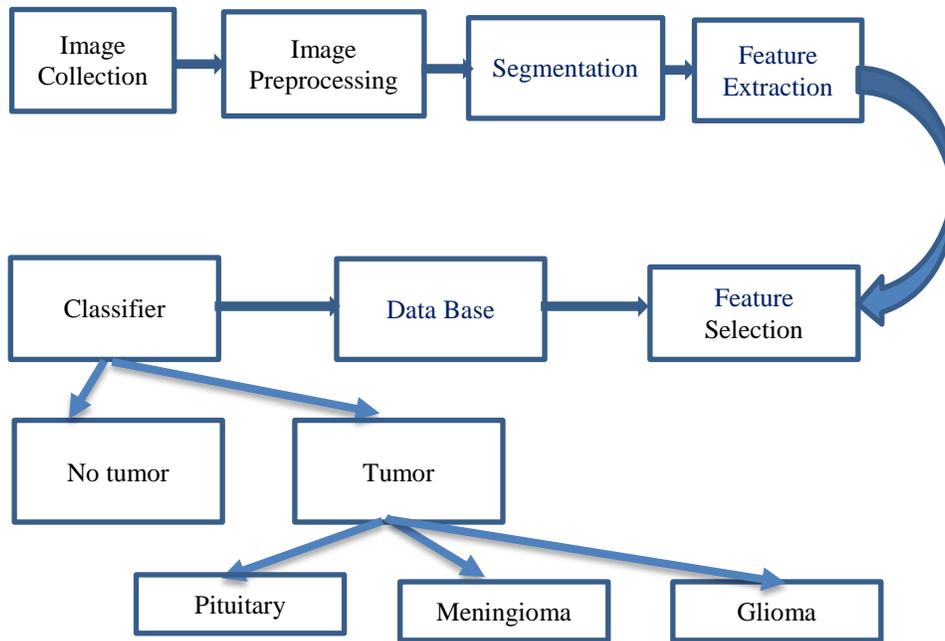


Fig. 1. Basic block diagram of proposed work

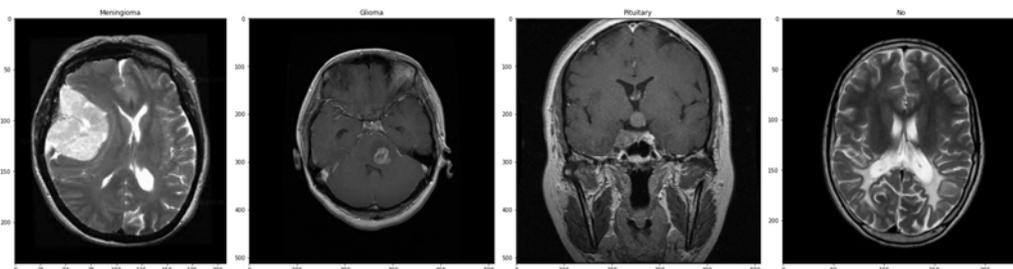


Fig. 2. (a) Meningioma (b) Glioma (c) Pituitary (d) No tumor

Our suggested method takes an exhaustive set of images as input and scales them all to the same dimensions as  $256*256*3$ . The input layer is typically administered by sixty-four convolutional filters of size  $2*2$ , and we use this to design a convolutional kernel that is both complex and efficient. Here we are using the ReLU activation function. If the input is positive, ReLU may be a piecewise linear operation that returns that value directly; otherwise, it returns zero. The pooling process involves summing the features that fall inside a 2D filter's coverage zone, as if the filter were dragged over each channel of the feature map. In a typical CNN model design, numerous convolutional and pooling layers are stacked on top of one another. Pooling layers is used to reduce dimensionality when working with large feature maps. As a result, the number of parameters that need to be learned to determine the level of computation carried out in the network is reduced. The pooling layer is responsible for summing up the features that are available in a certain portion of the feature map that was produced by a convolutional layer. Because of this, subsequent operations are carried out on summarized data

rather than correctly positioned features that were generated by the convolutional layer. Because of this, the model is more capable of distinguishing between distinct positions held by features in the input image.

Pooled feature maps are created after pooling. After merging many images into one, we need to flatten the resulting matrix into a column vector for further processing. After that, the data is sent to a neural network for evaluation. The use of two stacked, interconnected layers, the dense layer, was exemplified by Dense-1 and Dense-2. Keras processes the neural network using the dense function, and the resultant vector feeds the network's layer.

The RMS optimizer was utilised in the construction of the model, and binary cross-entropy was chosen to serve as the loss function. Because of this, we were able to evaluate the accuracy of the tumour detection. A suggested method for locating tumours is depicted in Fig. 3, and it makes use of a convolutional neural network (CNN).

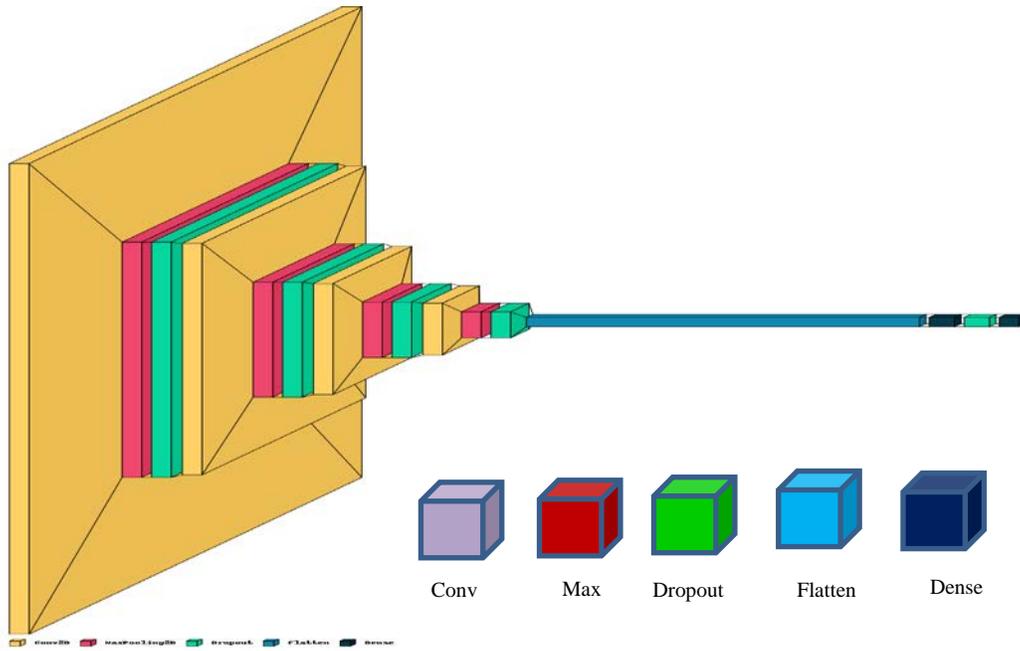


Fig. 3. Proposed methodology for tumor detection convolutional neural network

### III. RESULTS AND DISCUSSION

The effectiveness of the proposed network will be assessed based on the following quality metrics.

$$Sensitivity = \frac{TP}{TP+FN} * 100\% \quad (3)$$

$$Specificity = \frac{TN}{TN+FP} * 100\% \quad (4)$$

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} * 100\% \quad (5)$$

$$Precision = \frac{TP}{TP+FP} * 100\% \quad (6)$$

TABLE I. PERFORMANCE OF EXISTING AND PROPOSED ALGORITHMS

CNN Architecture	Sensitivity	Specificity	Precision	Accuracy
AlexNet	84.3	92.3	93.1	87.9
GoogleNet	84.8	95	95.5	88.6
VGG-16	81.2	87.3	88.1	83.4
Seetha et al. [13]	85	85.3	88.5	97.5
Tonmoy Hossain et al. [14]	86.2	91.5	93.5	97.87
Proposed Method	90	92	94.1	98.6

The comparison of several quality metrics with both existing and suggested methods is presented in Table I.

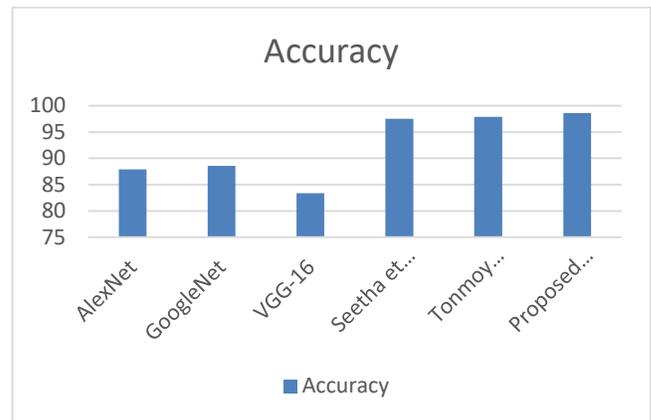


Fig. 4. Accuracy comparison

Fig. 4 shows the results of a comparison of different quality metrics using both existing and recommended methods. Fig. 5 depicts the model structure and various layers of the proposed model. Fig. 6 shows different MRI images- Prediction of brain tumours by the proposed model. Fig. 7 shows the accuracy and loss curves of the proposed model.

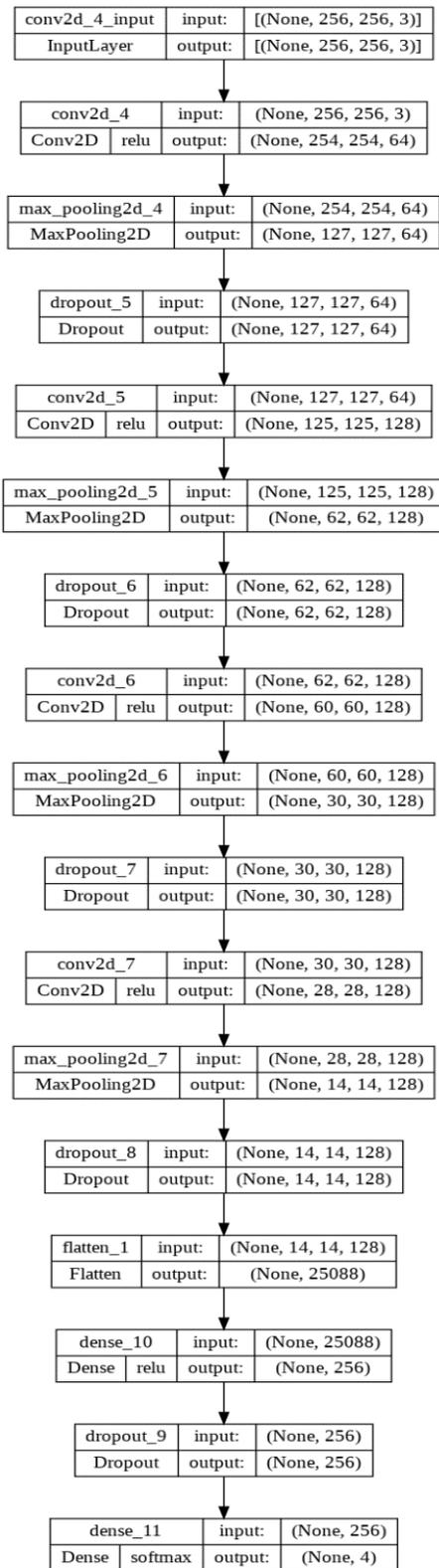


Fig. 5. Model structure and different layers of proposed model

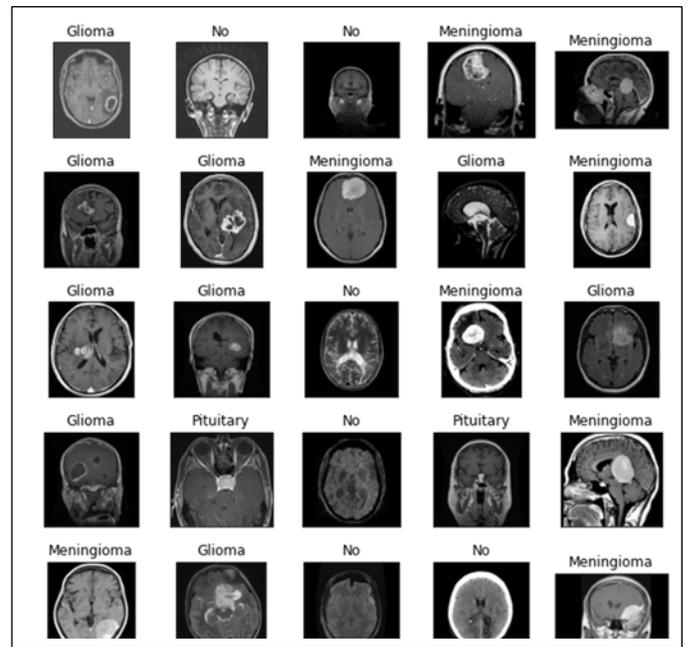


Fig. 6. Prediction of brain tumor by proposed model

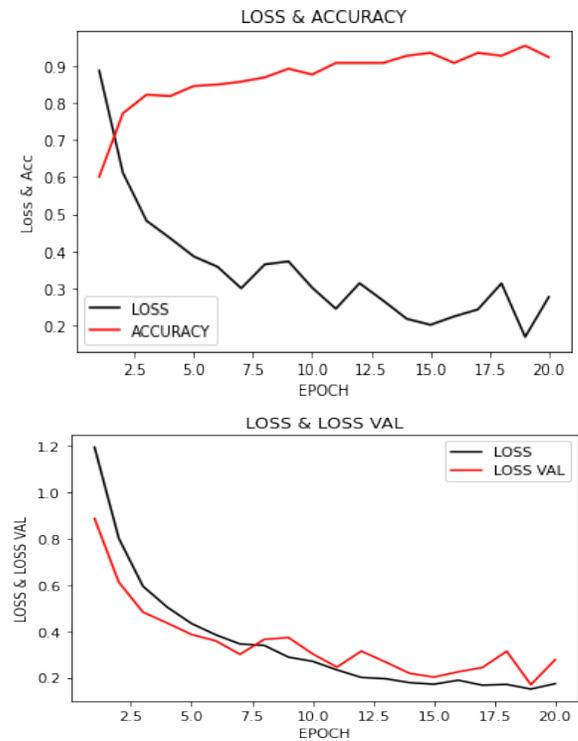


Fig. 7. The accuracy and loss curves of the proposed model

#### IV. CONCLUSION

Image segmentation is an essential part of medical image processing because of the extensive variety of medical images. The primary objective of this study is to develop a low-complexity, high-accuracy, automatic brain tumour categorization system. Images obtained from MRI and CT scans were analyzed and employed in the segmentation of the brain tumour. The proposed research was divided into four stages: data collection and pre-processing, data segmentation, MR image feature extraction, and data classification. In our research, we applied a method called global threshold segmentation for tumour segmentation, which accurately predicted the presence of tumour cells. The procedure of segmentation was then followed by classification using both conventional classifiers and a convolutional neural network. Following that, we put a variety of conventional classifiers to use and analyzed the results of each. The proposed method successfully distinguished between healthy and diseased tissues in MR images with an accuracy of 98.6%. The same method can also be utilised to detect and study a variety of disorders that are present in other areas of the body. Future research can improve accuracy by combining more effective segmentation and extraction methods with real-time images and clinical settings, as well as a large data set covering a wide range of conditions and classifiers with optimization methodology.

#### FUNDING DETAILS

No private, government, or non-profit organization provided direct money for his study.

#### INFORMED CONSENT

According to the author's declaration, there was no informed consent given.

#### DECLARATION OF COMPETING INTEREST

The authors have no conflicts of interest.

#### ETHICAL APPROVAL

This article does not contain any data or other information derived from experiments or research in which human or animal participants participated.

#### REFERENCES

- [1] John Prabha and J. Sathesh Kumar, "Performance evaluation of image segmentation using objective methods", *Indian Journal of Science and Technology*, Vol 9(8), February 2016.
- [2] D. Shen, G. Wu, H.-I. Suk, Deep learning in medical image analysis, *Annu. Rev. Biomed. Eng.* 19 (2017).
- [3] Yanming Sun, Chunyan Wang, A Computation efficient cnn system for high quality brain tumor segmentation, *Biomed. Signal Process Control* 74 (2022), 103475.
- [4] John Schmeelk, Wavelet transforms on two-dimensional images, 2002, pp. 939–948.
- [5] Carlos Parra, Robert Kozma, Automated brain data segmentation and pattern recognition using ann, 2003.
- [6] E.I. Papageorgiou, P.P. Spyridonos, Brain tumor characterization using the soft computing technique of fuzzy cognitive maps, *Appl. Soft Comput.* 8 (2008) 820–828.
- [7] El-Sayed Ahmed El-Dahshan, Tamer Hosny, Abdel-Badeeh M. Salem "Hybrid intelligent techniques for mri brain image classification" *ELSEVIER Digital Signal Processing*, vol. 20, 2010, pp. 433-441.
- [8] Fauzi Othman, Ariffanan Basri, Probabilistic neural network for brain tumor classification, 2011, pp. 136–138.
- [9] Pei L, Reza SMS, Li W, Davatzikos C, Iftekharruddin KM. "Improved brain tumor segmentation by utilizing tumor growth model in longitudinal brain MRI". *Proc SPIE Int Soc Opt Eng.* 2017.
- [10] Sajjad M, Khan S, Muhammad K, Wu W, Ullah A, Baik SW. Multi-grade brain tumor classification using deep cnn with extensive data augmentation. *J Comput Sci* 2019;30:174–82. doi:<https://doi.org/10.1016/j.jocs.2018.12.003>.
- [11] Kanmani P, Marikkannu P. MRI Brain Images Classification: A Multi-level threshold based region optimization technique. *J Med Syst* 2018;42:1–12. doi:10.1007/s10916-018-0915-8.
- [12] S. Sathesh, A. Sujatha Reddy, Noise suppression using wiener filtering in the non-sub-sampled contourlet domain for magnetic resonance brain images, *IEEE*, 2013
- [13] Seetha J, Raja S. S. Brain tumor classification using convolutional neural networks. *Biomed Pharmacol J* 2018;11(3). Available from: <http://biomedpharmajournal.org/?p=22844>
- [14] Tonmoy Hossain, Fairuz Shadmani Shishir, Mohsena Ashraf, M.D. Abdullah Al Nasim, Faisal Muhammad Shah, Brain tumour detection using convolutional neural network, in: 1st International Conference on Advances in Science, Engineering and Robotics Technology, ICASERT, 3-5 May 2019, 2019.
- [15] Srinivasarao Gajula, V. Rajesh (2020) A Review on recent trends in deep learning methods for medical image analysis. *International Journal of Psychosocial Rehabilitation*, 2020, Vol. 24, Issue 07, ISSN: 1475-7192.
- [16] Brain tumour: statistics, Cancer.Net Editorial Board, 1/2021. (Accessed on January 2021).
- [17] Yixin Wang, Yao Zhang, Feng Hou, Yang Liu, Jiang Tian, Cheng Zhong, Yang Zhang, Zhiqiang He Modality-pairing learning for brain tumour segmentation, arXiv:2010.09277v2 [eess.IV], 29 Dec 2020.
- [18] Gajula, S., Rajesh, V. An MRI brain tumour detection using logistic regression-based machine learning model. *Int J Syst Assur Eng Manag* (2022). <https://doi.org/10.1007/s13198-022-01680-8>
- [19] W. Zhang, Y. Wu, B. Yang, S. Hu, L. Wu, S. Dhelim, Overview of multi-modal brain tumour MR image segmentation, *Healthcare* 9 (2021) 1051, <https://doi.org/10.3390/healthcare9081051>.
- [20] Gupta, Gaurav and Vinay Singh. "Brain tumor segmentation and classification using FCM and support vector machine." (2017).
- [21] Xie, Xiaozhen A K-Nearest neighbor technique for brain tumor segmentation using minkowski distance. *J. Med.Imaging Health Inform.* 2018, 8, 180–185. [CrossRef].
- [22] Srinivasarao Gajula, V. Rajesh, 2021. "MRI brain image segmentation by using a deep spectrum image translation network". *Jour. of Med. P'ceutical & Alli. Sci.* V 10 - I 4, 1103, P-3097-3100. doi: 10.22270/jmpas.V10I4.1103.

# Classification of Psychological Disorders by Feature Ranking and Fusion using Gradient Boosting

## Classification of Psychological Disorders

Saba Tahseen<sup>1</sup>, Ajit Danti<sup>2</sup>

Ph.D. Scholar<sup>1</sup>, Professor<sup>2</sup>

Department of Computer Science and Engineering, Christ University,  
Bangalore, India<sup>2</sup>

**Abstract**—Negative emotional regulation is a defining element of psychological disorders. Our goal was to create a machine-learning model to classify psychological disorders based on negative emotions. EEG brainwave dataset displaying positive, negative, and neutral emotions. However, negative emotions are responsible for psychological health. In this paper, research focused solely on negative emotional state characteristics for which the divide-and-conquer approach has been applied to the feature extraction process. Features are grouped into four equal subsets and feature selection has been done for each subset by feature ranking approach based on their feature importance determined by the Random Forest-Recursive Feature Elimination with Cross-validation (RF-RFECV) method. After feature ranking, the fusion of the feature subset is employed to obtain a new potential dataset. 10-fold cross-validation is performed with a grid search created using a set of predetermined model parameters that are important to achieving the greatest possible accuracy. Experimental results demonstrated that the proposed model has achieved 97.71% accuracy in predicting psychological disorders.

**Keywords**—Electroencephalograph (EEG); psychological disorders; negative state emotions; gridSearchCV; gradient boosting classifier

### I. INTRODUCTION

A clinically severe issue with the capacity of an individual to think straight or rule out their emotions or behavior characterizes a psychological disorder. It is typically accompanied by anxiety or impairment in critical areas of functioning. There are many different sorts of mental diseases. Psychological problems are another term for mental problems. The latter is a larger phrase that includes mental disorders, psychological disabilities, and (additional) states of mind that cause severe distress, functional impairment, or the risk of self-harm. This preliminary report is about psychiatric illnesses. The 11th Revised Version of the Disease Classification System (ICD-11) estimates that 960 million people, or one out of every eight people, will live on the planet in 2019 and suffer from a mental disorder, with anxiety and depression being the most common. Depression disorders harmed 280 million people in 2019, including 23 million children and adolescents. Anxiety disorders afflict 301 million people, with 58 million of them being children and adolescents. Schizophrenia affects roughly 24 billion people worldwide or one in every 300. In 2019, 40 million people

were impacted by bipolar disorder [1]. As a result of the COVID-19 disease outbreak, the number of individuals living with anxiety and depression increased significantly in 2020. Estimations reveal a 26% spike in anxiety disorders and a 28% spike in serious depression disorders in only one year [2]. Many patients with psychiatric problems don't have access to suitable treatment or preventative options. Many people face stereotypes, marginalization, and infractions of their human rights. EEG is inexpensive and useful for assessing resting-state activity in the brain in natural environments, allowing for large amounts of data to be collected quickly. Furthermore, as the acquisition of technology improves and calculations improve, EEG is gathering steam as a foundational technology for brain-computer interfaces [3]. Relatively low cost, ease of use, and adaptable territory setup Echocardiography has been widely used in uncovering the aetiologies of various mental illnesses (e.g., depression [4], Alzheimer's [5], epilepsy [6], schizophrenia [7], autism spectrum disorder [8], anxiety [9], and so on). Typical brain activity and emotional swings are always present in depression, a mental disorder with clinical signs like severe depression and impaired thinking. EEG can therefore identify these aberrant events as a technique for monitoring brain activity.

Metadata and computational scientific research advances are being made in transforming mental healthcare. The scope of evidence that can be measured in terms of neural mechanisms and objective markers has expanded. Furthermore, the use of machine learning, also known as artificial intelligence, has grown. Machine learning can evaluate the effectiveness of forecasts on previously unseen (test) data that wasn't used prior, fitting the model for training data, utilizing out-of-sample forecasts, and providing individualized, and possibly high levels of therapeutic applications [10]. Machine learning is expected to aid or even replace physician judgments such as diagnostic testing, prognosis, and patient experience [11]. Individually, Cross-validation learning and psychological disorders had previously been investigated. Collaborations between these two fields have recently been combined, and machine learning has been used by researchers to identify psychological disorders using EEG data. Negative emotions like anxiety and depression encourage a series of psychological and physiological changes that put one's long-term health at risk. So for this, we sought to develop a new classification model for considering negative emotional features from EEG brainwave emotion features

(positive, negative, or neutral) in patients with severe psychological disorders. The contributions to this study are as follows:

- It is difficult but also necessary to accurately extract EEG features since the success of the classifying depends on this extraction. We extract negative state features from EEG brainwave data from emotional features (positive, negative, neutral) for psychological disorders.
- For negative state features we have applied the “Divide and Conquer” approach to four equal parts for finding more optimal results from each feature subset.
- We present the K-Means cluster technique to obtain labeled EEG signal features for each feature subset.
- Ensemble methods (RF-RFECV) techniques for feature selection have been employed to build and determine the highest-ranking features from each feature subset and merge them into an appropriate feature set for classifying psychological disorders.
- To obtain an optimal hyperparameter of the Gradient boosting, the tree classifier GridSearchCV has been used.
- Performance of hyperparameters of gradient boosting trees such as the number of trees, tree depth, several learning rates, and the number of subsamples are utilized.

This article is broken up into five sections. Beginning with the introduction. Most related works were described in Section II. Additionally, Section III provides materials and methods, and resources. In Section IV, Experimental Results are discussed. Finally, the conclusions are given in Section V.

## II. RELATED WORK

Most of the recent imaging research (i.e., employing magnetic resonance) has relied on supervised machine learning. To differentiate patients from health controls (HCs). Studies have primarily concentrated on Alzheimer's disease, schizophrenia, and depression, but have rapidly been expanded to include other diagnosing topics [12]. Literary work suggests that certain machine learning in EEG might predict significant psychological disorders and serve as an unbiased index of psychological disorders, according to the findings. The comprehension of the support vector machine, elastic net, and random forest machine learning methods was highlighted. The elastic net model with intelligence quotient adjustment performed the best with different bands of the EEG dataset [13]. According to various cross-validation experiments, the probability of diagnosis achieved employing the provided approach in a training dataset among 207 entities, such as 64 patients with severe depression, 40 patients with severe schizophrenia, 12 bipolar depression patients, and 91 healthy or normal subjects, is over 85% with some further advances [14]. The proposed methodology has the possibility of being a beneficial adjunct treatment tool for healthcare practitioners. EEG acquisition and pre-processing were adequate, discovered that numerous of them lacked thorough

clinical characteristic identification. Moreover, numerous studies employed parameters of the model or testing techniques that were flawed. Indeed, it's suggested that future researchers of psychological disorders using Deep Learning enhance the accuracy of clinical evidence and use cutting-edge model choice and test procedures to improve research standards and progress toward diagnostic value[15]. Using techniques for extracting multivariate EEG features and algorithms, the goal of this research is to provide an analytical framework for robotic GAD identification. Resting-state EEG data were collected from 45 GAD patients and 36 health control (HC) with 97.83 percent accuracy aberrated features helped classification performance [16]. This study classifies the EEGs of 43 VHS and 53 MDD patients using data mining techniques. This includes cleaning and normalizing the data beforehand, using Linear Discriminant Analysis (LDA) to map information into a brand-new feature space, and Genetic Algorithm (GA) to determine the much more features [17]. Methods are used in mental health to predict the likelihood of mental diseases and, thus, to execute prospective treatment outcomes. This review article lists many machine-learning techniques for identifying and diagnosing depression. The three classes of ML-based depression detection methods include deep learning, classification, and ensemble. The authors describe a generic paradigm for diagnosing depression that includes raw data, which was before, ML training; exposure, detection classifications, and performance assessment are all part of the process [18].

In [19], six channels (FT7, FT8, T7, T8, TP7, and TP8) are used to extract features from the frontal area of the brain. The following band powers—delta, theta, alpha, beta, gamma1, and gamma2—along with their related asymmetry and paired asymmetry—are employed as linear characteristics. The classifiers used are bagging and three different kernel functions of the Support Vector Machine (SVM) (polynomial, gaussian, and sigmoidal). Relief is the creation of predictive models utilizing the Decision Tree (DT) technique to find rules and relevant features; the feature selection method is applied. The feature selection methods SVM (Gaussian Kernel Function) and Relief were employed to achieve the best classification accuracy of 96.02% and 79.19% for the identification and severity rating of depression, respectively [19]. A widespread EEG system with three electrodes in the prefrontal lobe was used to record all electroencephalogram (EEG) signals from subjects during the sound stimulus and resting state at just the Fp1, Fp2, and Fpz electrolytic positions. A maximum of 270 linear and nonlinear features were recovered after denoising with the Finite Moment Generating Filter, which incorporates the Kalman derivation method, Discrete Wavelet Transformation, and an Adaptive Predictor Filter. The dimensionality of the feature space was then decreased using the minimal-redundancy-maximum-relevance feature extraction strategy. The depressed individuals were separated from the healthy controls using four classification techniques (SVM, KNN, Classification Trees, and ANN) [20].

The spatial frequency data of a few chosen EEG channels is used to extract features. Theta and beta bands were chosen as EEG frequency bands for this investigation using a

technique called "choosing a frequency range". The characteristics of the chosen frequency ranges of EEG are subject to feature selection. As for limitations of existing methods, it may be that they struggle with handling large datasets, lack effective feature selection algorithms, or have difficulty generalizing to new data. The proposed method may be designed to address these limitations and provide a more effective solution to the problem at hand. Finally, a variety of machine learning methods were used to categorize the chosen subset of characteristics from the statistically relevant EEG networks' proper frequencies. A random forest classification model with either nine or ten attributes is used. It is possible to classify anxiety on two or four levels with an accuracy of 94.90% and 92.74%, respectively [21]. The carefully chosen main studies were used in comprehensive mapping research. The objectives were to present a comprehensive picture of the most important research areas in the diagnosis and forecast of mental diseases by combining EEG with DL. [22]. SVM was used to categorize the stress levels using the labeled data from the k-means clustering method. Using only the beta-band ultimate power feature in the right (Fp2) prefrontal region, the achievement of the classification model was endorsed using the 10-fold cross-validation method. This result confirmed the excellent efficiency of 98% accuracy because of the significant adjustments in beta activity all through pre- and post-stimuli latent patterns using localized and reduced features and evaluating model accuracy and false positive findings on EEG data from people with MDD and HV. The motivation to write this research to solve the following issue came from considering the above kinds of literature and using abilities in this field [23]. Most of the research for psychological disorders has been done with resting states, eye open and closed states of EEG dataset.

- To find the best solution for a problem, it is important to consider all possible options and evaluate them based on relevant criteria. This involves a systematic and analytical approach to identifying the optimal solution.
- How can unsupervised learning data be effectively handled or processed, given the lack of labeled information, to improve the quality of the resulting models or insights?
- How to solve the overfitting issue in machine learning refers to identifying and implementing techniques that can prevent a model from becoming too complex and fitting too closely to the training data, which can result in poor performance on new, unseen data.
- Which parameter is used to develop the best model?
- Lack of standardization: Many studies in this field use different EEG acquisition protocols, pre-processing methods, feature extraction techniques, and machine learning algorithms, which make it challenging to compare results and generalize findings.

- Limited diagnostic focus: While some studies have investigated a range of mental health conditions, others have primarily focused on a few specific disorders, such as depression, anxiety, schizophrenia, or Alzheimer's disease. Further research is needed to evaluate the usefulness of machine learning with EEG for diagnosing other psychological disorders.

### III. PROPOSED METHODOLOGY

Psychological disorders are complex and can manifest in many ways, making diagnosis challenging for healthcare professionals. Machine learning algorithms have the potential to improve the accuracy and efficiency of diagnosis by identifying patterns in large datasets that may be difficult for humans to detect. The use of feature ranking and fusion in combination with gradient boosting is a promising approach for improving the performance of machine learning algorithms in the context of psychological disorder classification. Feature ranking techniques can help identify the most relevant features for classification, while feature fusion can combine different sources of information to improve the overall accuracy of the algorithm. Therefore, the motivation behind this research is to explore how these techniques can be applied to improve the accuracy and efficiency of psychological disorder classification using machine learning. The goal is to develop a more effective diagnostic tool that can assist healthcare professionals in accurately identifying and treating psychological disorders.

In this work, we have comprehensively analyzed the positive, negative, and neutral states of the publicly accessible EEG brainwave Dataset. In this research, for the recognition of psychological disorders, we extract the negative state of this dataset. The proposed method for the classification of psychological disorders using feature ranking and fusion with gradient boosting is an appropriate solution for the problem due to its ability to handle large datasets, and improve generalization, and flexibility. The method uses a feature selection algorithm to select the most important features for classification, which can enhance the accuracy and efficiency of the model. The proposed method also employs techniques to prevent overfitting, such as grid search for hyperparameter tuning, which can enable the model to generalize better to new and unseen data. This section outlines the configuration for the classification procedure, as well as the methodology used to conduct the research, and discusses a potential strategy for minimizing features in an EEG analysis by establishing the RF-RFECV method. The proposed architecture diagram shows the steps of this research in Fig. 1.

#### A. Data Preprocessing

Divide and Conquer strategy has been applied to the negative state features of the EEG brainwave dataset to find an optimal solution to a problem. Four distinct feature sets were compared to see how changing the measurements would affect the results and which feature set performed better [25].

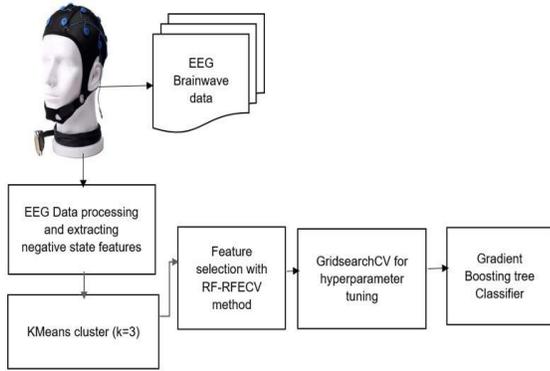


Fig. 1. Proposed architecture of psychological disorder classification

Steps of divide and conquer approach:

1) *Divide*: Divide the dataset  $D$  into 4 Feature subsets (FS)

$$D = \frac{(f_0, f_1, f_2, \dots, f_n)}{4} \quad (1)$$

$$= FS_1, FS_2, FS_3, FS_4$$

2) *Conquer*: Applied feature selection technique for each FS.

3) *Combine*: Selected features from each FS have been fusion into the final features set (FFS).

Further, the K-Means clustering technique [26] is applied to label unlabeled datasets to group the features into 3 similarity clusters.

The steps of K-Means are as follows:

- Initially, we generate random  $k$  points, referred to as means or cluster centroids.
- Every feature is categorized according to the nearest mean, and the precise location of that mean, which represents the average values of the features categorized in that cluster so far.
- Clusters are the result of repeating the process for a predetermined number of iterations.

The algorithm's final goal is to minimize the squared error function, which is represented by:

$$J(v) = \sum_{i=1}^c \sum_{j=1}^{c_i} (\|x_i - v_j\|)^2 \quad (2)$$

Where,

“ $\|x_i - v_j\|$ ” is a measure of how far the  $n$  data points are from each cluster's center [27]. In this research, the three clusters are chosen as 0, 1, and 2 for labeling EEG negative data as shown in Fig. 2

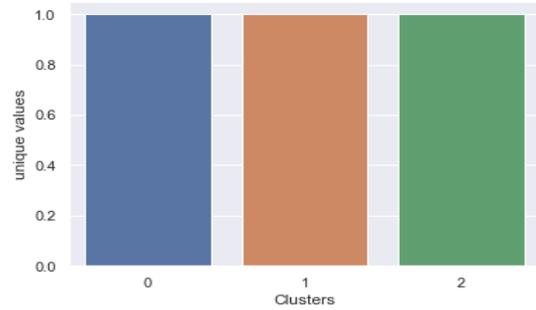


Fig. 2. Negative state features with 3 clusters i.e., 0, 1, 2

### B. Feature Selection

Selecting features from thousands of features is the most challenging research problem. RF-RFECV is used for feature selection. The algorithm is trained by Random Forest (RF) to generate the importance of features. The importance of each feature is calculated by equation 3.

$$fi_i = \frac{\sum_{j: \text{node } j \text{ split on feature } i} ni_j}{\sum_{k \in \text{all nodes}} ni_k} \quad (3)$$

RF classifier can be trained to produce feature importance values [28] that represent the relative importance of each feature. Following that, features are ranked in order of importance value. The component with the lowest importance value is eliminated. The classifier is then retrained using the remaining features until it runs out of features to train with. Finally, the complete ranking of the features can be obtained using the feature-importance-based RFE method i.e., (RF-RFECV). It has been demonstrated that, an embedded feature selection method performs well and makes up for the drawbacks of the filter and wrapper methods. The following pseudocode represents the proposed feature selection algorithm.

---

#### Algorithm 1: Feature Selection with RF-RFECV

---

Input: Feature Subset FS:  $\{f_0, f_1, f_2, \dots, f_n\}$   
 Output: Rank features according to smallest feature importance value,  $R$   
 Step 1: set  $R = \{\}$   
 Step 2: Repeat steps 3-9 until  $FS$  is not empty  
 Step 3: Train the RF using  $FS$ .  
 Step 4: Compute the importance of the feature with an equation (3).  
 Step 5: Determine the ranking method,  $\text{Rank} = fi^2$   
 Step 6: Rank the features in sorted orders.  
 $N_{rank} = \text{sort}(\text{Rank})$   
 Step 7: modification of the feature rank list  
 $\text{modify } R = R + FS(N_{rank})$   
 Step 8: Delete the features with the lowest rank  
 $\text{modify } FS = FS - FS(N_{rank})$   
 Step 9: Fusion the highest-rank features of  $FS_1, FS_2, FS_3, FS_4$  into the final dataset.  
 Return final dataset Fusion feature subset (FFS)

---

### C. Machine Learning Model

Gradient-boosting tree classifier is applied to predict the test results of psychological disorders. Gradient boosting is a sequenced method based on the ensemble principle. It integrates a group of weak learners and generates higher prediction accuracy. The model results are weighted based on the results of the initial instant  $t-1$  at any instant  $t$ . The working procedure gradient boosting is as follows.

Step 1: Create a fundamental model to predict the dataset. Take the total of the cluster column and presume that represents the expected value. The simple mathematical calculation behind these first steps.

$$F_0(x) = \arg \min_y \sum_{i=1}^n L(y_i, y) \quad (4)$$

Step 2: Determine the Residuals

$$e = (Y - \hat{Y}) \quad (5)$$

Step 3: Determine the decision tree's leaf output values

$$Y_m = \arg \min_Y \sum_{i=1}^n L(y_i, F_{m-1}(x_i) + Y h_m(x_i)) \quad (6)$$

Step 4: Update the prediction

$$F_m(x) = F_{m-1}(x) + V_m h_m(x) \quad (7)$$

### D. Hyperparameter Optimization by GridSearchCV

A method for identifying the best hyperparameters in a grid out of a set of parameters is called GridSearchCV. Gradient boosting trees could be a challenge to set up as an algorithm [29]. Grid search common ranges are produced because of the gradient boosting technique's key hyperparameters, which serve as the starting point for one's work. This might be done by assigning a dictionary for links, the names of the model hyperparameters, to the values to search for in the GridSearchCV [30]. The following is the procedure for finding the best hyperparameters with GridSeachCV.

#### Algorithm 2: GridSearchCV for hyperparameter optimization

**Input:** Dataset  $FFS = \{f_0, f_1, f_2, \dots, f_n\}$

**Output:** Best parameter with the highest accuracy

Step 1: Create the gradient boosting tree model and parameter

Step 2: Create a dictionary using the model's parameters.

- Develop an *estimator* of gradient boosting tree classifier
- Develop a *Param\_grid* with key hyperparameters  
The efficiency of model evaluation metrics.  
 $Score=f(Key Parameter)$
- Develop the *CV* for iterations

Step 3: Repeat the process step 2, going through each possible set of the grid's values one at a time.

Step 4: Fit the data set in the object function

Step 5: Run the objective functions multiple times per each possible pair of hyperparameter values.

**Return** the most accurate hyper-parameters available with the highest accuracy.

The GridSearchCV key hyperparameter for gradient-boosting trees is shown below in Table I.

TABLE I. KEY HYPERPARAMETERS FOR GRADIENT BOOSTING TREE WITH GRIDSEARCHCV

Model	n_estimators	Learning rate	subsample	Max depth
Gradient Boos tree	[10,50, 500,1000]	[0.0001, 0.001,0.01, 1.0]	[0.5, 0.7,1.0]	[0.5, 0.7,

### E. Performance Evaluations

Before the prediction model is constructed, a model must be assessed using several evaluation criteria [31]. To date, we have evaluated our prediction models using means and accuracy scores. However, the accuracy score and mean alone aren't always sufficient to assess a model adequately because it doesn't specify whether a class (positive or our models incorrectly forecast a negative) in the event of a poor accuracy rating this is clarified by precision score.

$$\text{Mean } \bar{x} = \left( \frac{\sum x}{n} \right) \quad (8)$$

$$\text{Std deviation SD} = \sqrt{\sum \left( \frac{x-\bar{x}}{n-1} \right)} \quad (9)$$

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (10)$$

$$\text{Precision} = \frac{TP}{TP+FP} \quad (11)$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad (12)$$

$$\text{score} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \quad (13)$$

$$\text{Confidence Score} = CI = \bar{x} \pm z \frac{s}{\sqrt{n}} \quad (14)$$

## IV. EXPERIMENTAL RESULTS

The study is significant in that it demonstrates the usefulness of an affordable and straightforward approach to diagnosing several psychological disorders using EEG. The proposed method employs a feature selection algorithm to select the most important features for classification, which can improve the accuracy and efficiency of the model. Additionally, the gradient boosting algorithm used in the method can handle large and complex datasets, making it well-suited for problems that involve many features and samples. The EEG brainwave dataset has been used, which has three states of emotion: positive, negative, and neutral. From these, three emotional states negative emotional states features were extracted for the recognition of psychological disorders for this work. 708 rows and 2548 columns of unlabeled data were used in the negative state. Applied the "Divide and Conquer" approach into 4 equal parts  $FS_1, FS_2, FS_3, FS_4$  to find more accurate results from each feature subset. The ensemble approach has been used for feature selection, followed by random forest feature importance with recursive feature elimination with cross-validation technique. Based on this procedure, selected the rank features for each subset, and fusion these new features into a new feature subset or new dataset.

An applied new dataset for the classification process used a gradient-boosting tree with GridSearchCV hyperparameter tuning. The best parameter combination is kept after the GridSearchCV evaluates all potential parameter value

combinations. This research primarily focuses on four parameters. GridSearchCV uses *max\_depth* for denoting the number of depths of a tree, *n\_estimators* for several sequential trees *learning\_rate* is used to determine how each tree will affect the predictions and, the *sub\_sample* for several analyses that will be chosen for every tree for the strongest impact on prediction accuracy. Several performance metrics are used for choosing the best parameter such as mean score, standard deviation score as well as accuracy.

The configuration that showed the best performance, achieving a means score of approximately 94.6 %, had a *learning\_rate* of 0.1, a *max\_depth* of 9 levels, 1000 *n\_estimators*, and a *sub\_sample* of 50% has presented in Table II.

Table II also presents the performance of parameters with different combinations of parameter values like *learning\_rate* of 1.0, 1000 of *n\_estimators*, *Subsample* of 40%, and *max\_depth* of 7 achieved means score is 92%.

*learning\_rate* of 0.2, 500 of *n\_estimators*, *Subsample* of 90% and *max\_depth* of 6 achieved means score is 91%.

*learning\_rate* of 1.0, 1000 of *n\_estimators*, *Subsample* of 90% and *max\_depth* of 7 achieved means score is 89%.

Accuracy of each parameter combination 91%, 89%, 92%, 92%, 96.71%.

TABLE II. BEST PARAMETERS FROM THE GRIDSEARCHCV METHOD

Sl.no	Best parameter	Means Score	Std score	Accuracy
1	n_estimators=1000, Subsample=0.4, learning_rate=1.0, max_depth=7	92.00%	0.028	91.00%
2	n_estimators=500, Subsample=0.9, learning_rate=0.2, max_depth=6	91.00%	0.26	89.00%
3	n_estimators=1000, Subsample=0.9, learning_rate=1.0, max_depth=7	92.00%	0.028	92.00%
4	n_estimators=1000, Subsample=0.9, learning_rate=1.0, max_depth=7	89.00%	0.24	92.00%
5	n_estimators=1000, subsample=0.5, learning_rate=0.1, max_depth=9	94.93%	0.27	96.71%

Additionally, the accuracy of each cluster is 0, 1, 2, and has achieved 95.5%, 96.8%, and 100% accuracy. The classification results in representation in Table III for clusters of 0, 1 a, and 2, Precision, Recall, and f1-Score achieved the highest accuracy in cluster 2 compared to other clusters for the proposed gradient boosting Classifier with GridSearchCV for a new dataset.

TABLE III. CLASSIFICATION REPORT OF EACH CLUSTER

Clusters	precision	recall	F1 score
0	94%	96%	95%
1	98%	97%	97%
2	100%	100%	100%

Fig. 3 represents the accuracy analysis of four algorithms with the x-axis being proposed and existing algorithms and the y-axis being proposed as the accuracy value. The accuracy of the proposed classifier without ranked features is 94% and with ranked features is 96.71%.

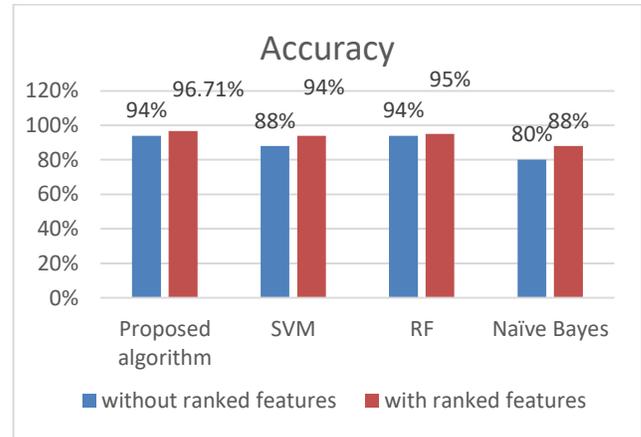


Fig. 3. Comparative analysis with another existing algorithm

The accuracy of SVM without ranked features is 92% and with ranked features is 94%. The accuracy of RF with ranked features is 95% and without ranked features is 94%. The accuracy of Naïve Bayes classifier achieved 80% accuracy without ranked features and 88% accuracy with ranked features. It can be concluded that the proposed algorithm with ranked features (new dataset) produces effective results as compared to existing algorithms.

Table IV displays the Confidence score of various psychological disorders (stress, depression, bipolar disorder, anxiety, autism, schizophrenia, mood disorder, and personality disorder) in an EEG negative emotional state with clusters 0, 1, and 2, which is represented. Schizophrenia disorders achieved an 85% confidence score as compared to other disorders.

TABLE IV. THE CONFIDENCE SCORE OF PSYCHOLOGICAL DISORDER

Psychological disorder	Accuracy	Negative Emotional State (0,1,2)
Depression	65.00	Cluster-0
Anxiety	35.00	Cluster -0
Stress	77.00	Cluster -0
Bipolar Disorder	50.50	Cluster -1
Personality	58.00	Cluster -1
Schizophrenia	85.00	Cluster -2
Autism	50.00	Cluster-2

Table V compares the state-of-the-art methods of other machine learning methods with the same dataset our proposed algorithm achieved the highest accuracy.

TABLE V. COMPARATIVE ANALYSIS OF PROPOSED WORK WITH OTHER WORK

Study	Classifiers	Datasets	Accuracy
The proposed method	Gradient Boosting Tree with GridSearchCV	EEG brainwave dataset	96.71%
[24]	Adaptive Boosted LSTM and DevoMLP	EEG brainwave dataset	85%
[31]	RNN	EEG brainwave dataset	95%
[32]	XGBoost	EEG brainwave Dataset	95%

## V. CONCLUSION

The study has demonstrated the usefulness of an affordable and straightforward approach to the brain utilizing EEG for the diagnosis of several psychological disorders, including stress, bipolar disorder, autism, mood, personality, anxiety, and depression. One significant advantage of the proposed method is its ability to handle large and complex datasets using gradient boosting, which is a powerful algorithm for handling such data. In addition, the techniques used to prevent overfitting, such as grid search for hyperparameter tuning, can help the model generalize better to new and unseen data. This is an important consideration when dealing with medical data where the model's ability to generalize to new data is crucial for accurate diagnosis. This study presents a unique method of feature selection with different feature subsets and makes a new dataset with 1300 features with the RF-RFECV algorithm with labeling using the K-Means Cluster technique. To address overfitting and optimize the parameters of the gradient boosting classifier, we employed the GridSearchCV algorithm to find the optimal hyperparameters for predicting psychological disorders from the EEG brainwave dataset, which has achieved 96.71% accuracy in classifying psychological disorders using negative states of emotion. The future will be, to calculate the severity of the psychological disorder and develop a web application for clinical diagnostics.

## ACKNOWLEDGMENT

None of the funding sources for this study were public, private, or nonprofit. We sincerely thank the professors of Christ University's department of computer science and engineering, Bangalore, India.

## CONFLICT OF INTEREST

The authors declare no conflict of interest.

## REFERENCES

- [1] Institute of Health Metrics and Evaluation. Global Health Data Exchange (GHDx), (<https://vizhub.healthdata.org/gbd-results/>, accessed 14 May 2022).
- [2] Mental Health and COVID-19: Early evidence of the pandemic's impact. Geneva: World Health Organization; 2022.
- [3] G. Li, B. Li, Y. Jiang, W. Jiao, H. Lan, and C. Zhu, "A new method for automatically modeling brain functional networks," *Biomedical Signal Processing and Control*, vol. 45, pp. 70–79, Aug. 2018, doi: 10.1016/j.bspc.2018.05.024.

- [4] A. Khalid, B. S. Kim, M. K. Chung, J. C. Ye, and D. Jeon, "Tracing the evolution of multi-scale functional networks in a mouse model of depression using persistent brain network homology," *NeuroImage*, vol. 101, pp. 351–363, 2014.
- [5] M. Cejnek, O. Vysata, M. Valis, and I. Bukovsky, "Novelty detection-based approach for alzheimer's disease and mild cognitive impairment diagnosis from EEG," *Medical & Biological Engineering & Computing*, vol. 59, no. 11-12, pp. 2287–2296, 2021.
- [6] S. C. Ponten, F. Bartolomei, and C. J. Stam, "Small-world networks and epilepsy: Graph theoretical analysis of intracerebrally recorded mesial temporal lobe seizures," *Clinical Neurophysiology*, vol. 118, no. 4, pp. 918–927, 2007.
- [7] S. Micheloyannis, E. Pachou, C. J. Stam, M. Breakspear, P. Bitsios, M. Vourkas, S. Erimaki, and M. Zervakis, "Small-world networks and disturbed functional connectivity in schizophrenia," *Schizophrenia Research*, vol. 87, no. 1-3, pp. 60–66, 2006.
- [8] S. Carlier, S. Van der Paelt, F. Ongenaes, F. De Backere, and F. De Turck, "Empowering children with ASD and their parents: Design of a serious game for anxiety and stress reduction," *Sensors*, vol. 20, no. 4, p. 966, 2020.
- [9] H. Byeon, "Exploring factors for predicting anxiety disorders of the elderly living alone in South Korea using Interpretable Machine Learning: A population-based study," *International Journal of Environmental Research and Public Health*, vol. 18, no. 14, p. 7625, 2021.
- [10] Q. J. Huys, T. V. Maia, and M. J. Frank, "Computational psychiatry as a bridge from neuroscience to clinical applications," *Nature Neuroscience*, vol. 19, no. 3, pp. 404–413, 2016.
- [11] G. E. Simon, "Big data from Health Records in Mental Health Care," *JAMA Psychiatry*, vol. 76, no. 4, p. 349, 2019.
- [12] D. B. Dwyer, P. Falkai, and N. Koutsouleris, "Machine learning approaches for clinical psychology and psychiatry," *Annual Review of Clinical Psychology*, vol. 14, no. 1, pp. 91–118, 2018.
- [13] S. M. Park, B. Jeong, D. Y. Oh, C.-H. Choi, H. Y. Jung, J.-Y. Lee, D. Lee, and J.-S. Choi, "Identification of major psychiatric disorders from resting-state electroencephalography using a machine learning approach," *Frontiers in Psychiatry*, vol. 12, 2021.
- [14] A. Khodayari-Rostamabad, J. P. Reilly, G. Hasey, H. de Bruin, and D. MacCrimmon, "Diagnosis of psychiatric disorders using EEG data and employing a statistical decision model," *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology*, 2010.
- [15] M. de Bardeci, C. T. Ip, and S. Olbrich, "Deep learning applied to electroencephalogram data in mental disorders: A systematic review," *Biological Psychology*, vol. 162, p. 108117, 2021.
- [16] Z. Shen, G. Li, J. Fang, H. Zhong, J. Wang, Y. Sun, and X. Shen, "Aberrated multidimensional EEG characteristics in patients with generalized anxiety disorder: A machine-learning based analysis framework," *Sensors*, vol. 22, no. 14, p. 5420, 2022.
- [17] M. Mohammadi, F. Al-Azab, B. Raahemi, G. Richards, N. Jaworska, D. Smith, S. de la Salle, P. Blier, and V. Knott, "Data mining EEG signals in depression for their diagnostic value," *BMC Medical Informatics and Decision Making*, vol. 15, no. 1, 2015.
- [18] S. Aleem, N. ul Huda, R. Amin, S. Khalid, S. S. Alshamrani, and A. Alshehri, "Machine learning algorithms for depression: Diagnosis, insights, and Research Directions," *Electronics*, vol. 11, no. 7, p. 1111, 2022.
- [19] S. Mahato, N. Goyal, D. Ram, and S. Paul, "Detection of depression and scaling of severity using six channels EEG Data," *Journal of Medical Systems*, vol. 44, no. 7, 2020.
- [20] H. Cai, J. Han, Y. Chen, X. Sha, Z. Wang, B. Hu, J. Yang, L. Feng, Z. Ding, Y. Chen, and J. Gutknecht, "A pervasive approach to EEG-based depression detection," *Complexity*, vol. 2018, pp. 1–13, 2018.
- [21] F. Muhammad and S. Al-Ahmadi, "Human State Anxiety Classification framework using EEG signals in response to exposure therapy," *PLOS ONE*, vol. 17, no. 3, 2022.
- [22] M. J. Rivera, M. A. Teruel, A. Maté, and J. Trujillo, "Diagnosis and prognosis of mental disorders by means of EEG and Deep Learning: A Systematic Mapping Study," *Artificial Intelligence Review*, vol. 55, no. 2, pp. 1209–1251, 2021.

- [23] T. Y. Wen and S. A. Mohd Aris, "Hybrid approach of EEG stress level classification using K-means clustering and support vector machine," *IEEE Access*, vol. 10, pp. 18370–18379, 2022.
- [24] J. J. Bird, D. R. Faria, L. J. Manso, A. Ekárt, and C. D. Buckingham, "A deep evolutionary approach to bioinspired classifier optimisation for brain-machine interaction," *Complexity*, vol. 2019, pp. 1–14, 2019.
- [25] S. E. Sánchez-Hernández, R. A. Salido-Ruiz, S. Torres-Ramos, and I. Román-Godínez, "Evaluation of feature selection methods for classification of epileptic seizure EEG signals," *Sensors*, vol. 22, no. 8, p. 3066, 2022.
- [26] A. Bablani, D. R. Edla, V. Kuppili, and D. Ramesh, "A multi stage EEG data classification using K-means and feed forward neural network," *Clinical Epidemiology and Global Health*, vol. 8, no. 3, pp. 718–724, 2020.
- [27] A. Bi, W. Ying, and L. Zhao, "Fast enhanced exemplar-based clustering for incomplete EEG signals," *Computational and Mathematical Methods in Medicine*, vol. 2020, pp. 1–11, 2020.
- [28] H. AlSagri and M. Ykhlef, "Quantifying feature importance for detecting depression using Random Forest," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 5, 2020.
- [29] L. Hertel, P. Baldi, and D. L. Gillen, "Reproducible hyperparameter optimization," *Journal of Computational and Graphical Statistics*, vol. 31, no. 1, pp. 84–99, 2021.
- [30] Y. Xie, C. Zhu, W. Zhou, Z. Li, X. Liu, and M. Tu, "Evaluation of machine learning methods for formation lithology identification: A comparison of tuning processes and model performances," *Journal of Petroleum Science and Engineering*, vol. 160, pp. 182–193, 2018.
- [31] M. K. Chowdary, J. Anitha, and D. J. Hemanth, "Emotion Recognition from EEG Signals Using Recurrent Neural Networks," *Electronics*, vol. 11, no. 15, p. 2387, Jul. 2022, doi: 10.3390/electronics11152387.
- [32] S. Klibi, M. Mestiri and I. R. Farah, "Emotional behavior analysis based on EEG signal processing using Machine Learning: A case study," *2021 International Congress of Advanced Technology and Engineering (ICOTEN)*, Taiz, Yemen, 2021, pp. 1-7, doi: 10.1109/ICOTEN52080.2021.9493537.

# Public Response to the Legalization of The Criminal Code Bill with Twitter Data Sentiment Analysis

Deny Irawan<sup>1</sup>, Dana Indra Sensuse<sup>2</sup>, Prasetyo Adi Wibowo Putro<sup>3</sup>, Aji Prasetyo<sup>4</sup>  
Faculty of Computer Science, Universitas Indonesia, Jakarta, Indonesia<sup>1, 2, 3, 4</sup>

**Abstract**—The Criminal Code Bill, also known as Rancangan Kitab Undang-undang Hukum Pidana (RKUHP), passed in the House of Representatives (DPR) on December 6, 2022, is being debated because several issues need to be fixed. Therefore, research was conducted to determine the public's reaction to the ratification of the Criminal Code Bill by analyzing Twitter data. This study aims to obtain a general response to the legalized RKUHP. We use sentiment analysis, a text-processing method, to get data from the public. To do this, we used N-grams (unigrams, bigrams, and trigrams) along with three algorithms: Naïve Bayes, Classification and Regression Tree (CART), and Support Vector Machine (SVM). The result of sentiment analysis found that 51% of tweets were positive about the ratification of the RKUHP, and 49% were negative. In addition, it was also found that SVM has the best accuracy compared to other algorithms, with an accuracy value of 0.81 on the unigram combination.

**Keywords**—Sentiment analysis; RKUHP; support vector Machine (SVM); Naïve Bayes; classification and regression tree (CART)

## I. INTRODUCTION

The Criminal Code Bill, also known as RKUHP, was signed into law on December 6, 2022. This is a momentous occasion, as this law would replace the existing Criminal Code (KUHP) [1]. Because the current Criminal Code is a legacy of Dutch colonialism, it is a version of the *Wetboek van Strafrecht voor Nederlandsch-Indie*. Changes are needed because the old Criminal Code is not keeping up with the times [2]. In addition, the revision of the previous RKUHP was carried out in partly making new laws related to the KUHP, which made regulations run wild, had no system or pattern, was inconsistent, made problematic laws, and even damaged the basic building system old KUHP [3]. RKUHP will be valid for three years from the date of promulgation [4]. However, the RKUHP is considered to have problematic articles still. Nurina S. (2022), in an interview with The Guardian, stated that at least 88 reports contain broad provisions that can be exploited and misconstrued by both the government and the general public to punish anyone and suppress the freedom of expression [5].

The response to ratifying RKUHP still has pros and cons. Because responses are essential determinants of every human action, interesting to see the public's response to RKUHP; when deciding, we need others' opinions. Companies or governments must know how the public feels about their products and services. The public sometimes utilizes Facebook and Twitter to engage socially online. Web-based social networks gradually engage the public [6]. This is consistent with research conducted in the United States about the public's

reaction to the Chicago Department of Public Health's laws on electronic cigarettes, which examined the public's response on Twitter. The data can help organizations predict, recognize, and respond to how the community will react by finding patterns in how people have responded to this policy [7]. In addition, according to research conducted in Mexico, governments frequently use Twitter to interact with their citizens. As a result, it has emerged as a valuable source of information for studying how governments interact with their constituents and how those citizens respond to those communications. These insights about how people interact with the government can be used to help make public policies and understand how the public sees those policies [8].

The ratification of RKUHP has the same context. It will be fascinating to watch how Twitter data is utilized to gauge public opinion toward the ratification of RUU KUHP because these messages on Twitter are openly accessible. Consequently, it can be viewed as raw data primarily for the extraction of opinions and for the analysis of policy by analyzing the sentiment [9]. This will aid the government's ability to forecast, detect, and respond to the public's reaction to the dissemination of information before it is completely implemented. Sentiment analysis is another term for "opinion mining" or "emotion Artificial Intelligence". It refers to applying natural language processing (NLP), Using text mining, computational linguistics, and biometrics to carefully identify, extract, assess, and look into people's emotions and personal data [6].

This research aims to identify the sentiments surrounding the ratification of the RKUHP. The analysis results are reprocessed to determine what aspects of RKUHP concern the public. The use of public sentiment will assist the government in gauging the public's reaction to the ratification of the RKUHP and can be utilized as input for the planned socialization. In addition, by using multiple algorithm models, this research will identify the optimal categorization model that might be used by the government when trying to determine public responses with data from twitter.

This research consists of five sections. The introduction, which contains the research's context and objectives, is the first section. The second section is a review of previous research and the theoretical framework. The study's research methodology is described in the third section. The fourth section is the results and discussion, which includes the findings from the research. The conclusion is the concluding section of the study.

## II. LITERATURE REVIEW

### A. Previous Research

This section contains considerable research that employs various methods for sentiment analysis. The first research authors use one methodology for measurement sentiment analysis, shown in Table I. Authors in [10] have investigated the Naïve Bayes algorithm's capacity to classify public mood under COVID-19's new normal. From the 2807 tweets that have been processed, the test results show that Naïve Bayes has done an excellent job, with an accuracy of 83% and an F1-score of 84%. The author in [11] researched sentiment analysis using the Support Vector Machine (SVM) with Weka (Waikato Environment for Knowledge Analysis) method and tested it on three different data sets with various labels. Because of this, the data set with the highest f1-score is the third one, which only has two titles: positive and negative.

Further research uses two methodologies for measurement sentiment analysis. In a study [12], sentiment analysis tests on comments on YouTube using Naïve Bayes and Support Vector Machines (SVM). Results when using a data scale of 7:3, with 70% of the data used for training and 30% for testing, show that the combination of Naïve Bayes and SVM results in higher accuracy and superior performance. In a study [13], researchers compare Naïve Bayes and SVM to evaluate the classification results that each method produces. Twitter data is used in this study for Tokopedia services. The outcomes demonstrated that, with an accuracy of 83.34%, the SVM linear kernel technique surpassed the Naïve Bayes technique. In a study [14] Using Twitter data, researchers assess the sentiment analysis of the COVID-19 virus infection on Indonesian public transportation. In this study, the authors used two comparison methods: Naïve Bayes and decision trees. The result is that Naïve Bayes outperforms the Decision Tree with an accuracy of 73.59%.

The third research uses more than two methodologies for measurement sentiment analysis. In a study [15], Researchers researched the sentiment analysis of tourists in Thailand during the COVID-19 pandemic. This study used three methods: SVM, Classification and Regression Tree (CART), and random forest. Consequently, SVM could identify the attitudes and intentions of the English-language tweets that included Phuket and Chiang Mai the best. Still, for tweets mentioning Bangkok, CART is the most accurate, with accuracies of 94.3%. Bangkok has more data tweets than others. Subsequent research, customer reviews of Amazon products. Researchers in this study [16] used four sentiment analysis methods: Naïve Bayes, SVM, Decision Tree, and K-Nearest Neighbor. In addition, this research also added TF-IDF and N-gram to its processing. The results of the TF-IDF method with N-grams show unigrams with SVM were the maximum accuracy results for Amazon product customer reviews. This study also found that comments on Amazon products influence potential consumers' purchasing decisions. The two studies [15][16] were conducted to determine the differences and accuracy of the sentiment analysis method.

TABLE I. PREVIOUS RESEARCH

No	Ref	Algorithm and Method	Sentiment Analysis and Objective	Result
1	[10]	Naïve Bayes	Provides a sentiment analysis of how well society is accepting the new normal with data from Twitter and investigates the Naïve Bayes algorithm's capacity to classify public mood under COVID-19's new normal.	During the COVID-19 pandemic, the majority of people were able to adjust to their new everyday normal. The test results show that Naïve Bayes has done an excellent job.
2	[11]	SVM with Weka (Waikato Environment for Knowledge Analysis)	This study didn't specifically look at sentiment analysis. Instead, it used three different data sets, two from Twitter and one from the Internet Movie Database, to test how well the SVM algorithm worked (IMDB).	The highest-scoring data set is the third (IMDB data), with only two titles: positive and negative.
3	[12]	Naïve Bayes and SVM	Provides a sentiment analysis of positive and negative YouTube comments and evaluates the combination of two algorithms, naïve Bayes and SVM.	The combination of Naïve Bayes and SVM results in higher accuracy and superior performance for seeing sentiment in YouTube comments.
4	[13]	SVM and Naïve Bayes	Sentiment analysis for Tokopedia service with data from Twitter and evaluation of the performance of Naïve Bayes and SVM.	The data do not specify the sentiment analysis results for the Tokopedia service; they evaluate that SVM is more accurate than Naïve Bayes.
5	[14]	SVM and Decision Tree	Sentiment analysis to determine what commuter line riders think about how the Covid-19 pandemic could spread on public transportation—and Comparison Accuracy from algorithm naïve Bayes and decision tree.	Most people in the community have a positive outlook that includes a plea and a call to stop the COVID-19 outbreak and get it under control. With an accuracy of 73.59%, Naive Bayes is better than the Decision Tree.
6	[15]	decision tree, random forest, and SVM with TF-IDF and combination ngram (unigram, bigram, and trigram)	Sentiment analysis to find out the expression of tourists about tourist attractions, events, festivals, and experiences from July to December 2020 whit data from Twitter.	The results showed the top 10 words for each type of feeling, which can be looked at to learn more and give the right advice.
7	[16]	SVM with a combination of term weighting and ngram	This study aims to assess the impact of sentiment (positive, negative, and neutral) and Amazon product reviews on sales performance. Also, to	The Result found that comments on Amazon products influence potential consumers' purchasing decisions. In

			identify the optimal combination of SVM, TF-IDF, and ngram.	addition, the TF-IDF method with N-grams shows unigrams with SVM were the maximum accuracy results for Amazon product customer reviews.
--	--	--	---	---

Based on previous research, researchers will use the Naïve Bayes [10][13][14], SVM [11][12] [13][15][16], and CART [15] in evaluating sentiment analysis. In addition, the N-gram and TF-IDF methods will be used because they are proven to increase accuracy [16]. The study used positive and negative labels because it was established in research [11] that they have the highest accuracy compared to data using more than two labels.

### B. Sentiment Analysis

According to Pang et al. (2002), opinion mining and sentiment analysis are two terms that refer to the same process. Sentiment analysis automatically analyzes, extracts, and textually processes material to derive the sentiment information in a single opinion sentence. An individual's perspective, or their predisposition to have a positive or negative view or opinion about a particular issue or object, can be determined using a technique known as "sentiment analysis" [17] [12].

### C. Data Preprocessing

Data Preprocessing involves converting raw data into a format the user may understand. Frequently, the data must be more structured and consistent, lack specific behaviors or patterns, and contain missing values, all of which contribute to many errors. Consequently, it needs to be cleaned, integrated, altered and decreased. The noise is eliminated, and missing values are filled in when cleaning is performed [18][19].

### D. N-Gram

The word n-gram feature counts sets of sequential N words in each tweet, where N can range from 1 to N. [20]. N-grams can be more informative. There could be  $t^2$  bigrams containing t different words. In practice, only some characteristics are generated because terms can't follow each other. Usually, n-grams are more distinct than words. A more extensive, less common feature space is an n-gram. A larger n increases information and computational expense [21]. In this research, we combine the unigram, the bigram, and the trigram forms of the n-gram.

### E. Term Frequency - Inverse Document Frequency (TF-IDF)

According to Jones (1972), Inverse Document Frequency (IDF) is a technique that can be combined with term frequency to lessen the influence of implicitly famous words in the corpus. This is how IDF is meant to be used. IDF gives greater weight to terms that appear more frequently in the document,

regardless of whether those words are used often or infrequently [22][23]. TF-IDF is now the most popular text classification and document categorization scheme [24][21].

### F. Naïve Bayes Algorithm

This categorization method is based on Bayes' Theorem and makes strong (naive) assumptions about feature independence. A Naïve Bayes classifier makes the following assumptions: that the proximity of one feature (element) within a class is unrelated to the proximity of other items. The Naïve Bayes algorithm is often used to divide texts into different groups, and it was recently used to separate data from sentiment analysis into groups [6].

The algorithm relies on Bayes' theorem and presumes that the class variable's value provides information for all variables independently. It is simple to program the Naïve Bayes classification algorithm to perform exceptionally well in supervised learning, and it can also be used in difficult real-world situations. The Naïve Bayes method is simple to grasp, needs an education dataset to figure out how to calculate its variables, doesn't care about things that have nothing to do with the problem, and works well with correct data from a single source [25][10].

### G. Support Vector Machine (SVM) Algorithm

According to Han et al. (2012), the Support Vector Machine (SVM) algorithm's goal is to locate the Maximum Marginal Hyperplane (MMH) by utilizing margins and support vectors. The MMH hyperplane is the best one available since it has the most significant margin distance and can be used to accurately and maximally segregate data for each class. Suppose both margins are in a position that is parallel to the hyperplane. In that case, the margin is defined as the point at which the shortest distance from a hyperplane to one side equals the distance from the hyperplane to the other side of the margin [26][24].

### H. Classification and Regression Tree (CART)

The classification and regression trees (CART) method is a systematic technique that was developed by Breiman et al. (1984) [27][28]. For the construction of decision trees, CART employs historical data. The dependent variable decides whether a classification tree (for categorical categories) or a regression tree (for variables with continuous categories) will be formed. The newly discovered observations can then be predicted (using a regression tree) and classified (using a classification tree) using the constructed tree. Contrary to classification trees, regression trees do not have any pre-determined classes. On the other hand, classification trees allow the user to select or calculate dependent variable types based on an external criterion. [27][29][30][28]. The CART approach consists of three steps: (1) the creation of the entire tree; (2) the selection of the ideal tree size; and (3) the evaluation of the results. (3) using a built tree to organize data or generate new information[28].

### III. METHODOLOGY

The research consisted of several stages, including the collection of data, the creation of data sets, the labeling of data, the processing of data, the grouping of words using n-grams and term weighting (TF-IDF), classification modeling, the evaluation of classification modeling, and, finally, the output of sentiment results and recommendations. This is shown in **Error! Reference source not found.**

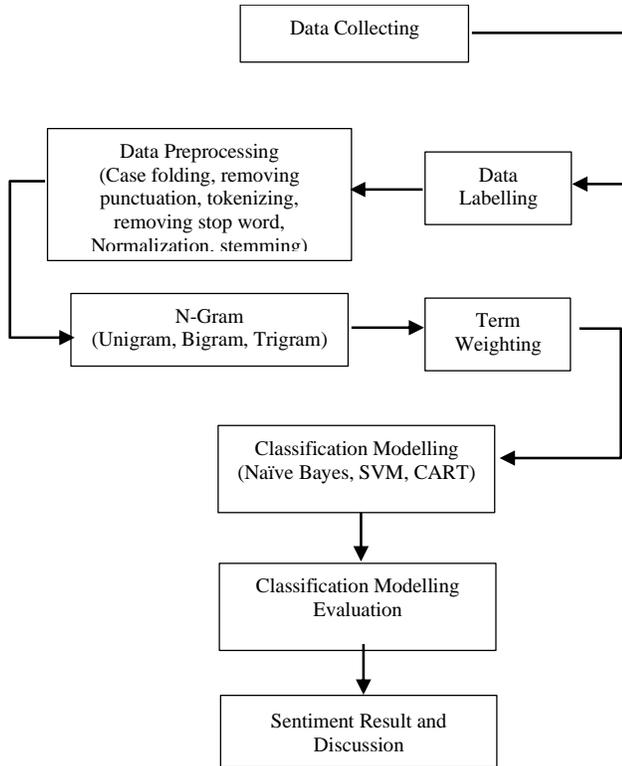


Fig. 1. Methodology for research

#### A. Data Collecting

Python and the twitter-snsrcape library package are used to harvest Twitter tweet data at this step. Data was gathered using a search for the phrase "RKUHP" tweeted between December 6, 2022, and December 31, 2022. Tweets taken are in Indonesian, and identical tweets will be deleted. Related Tweets that only use the RKHUP hashtag and only contain ads will also be disqualified. Tweets are not converted into English due to possible differences in meaning in processing. All words resulting from sentiment will use the Indonesian language.

#### B. Data Labelling

In this phase, the training data is labeled manually whether tweets are positive (pro) or negative (con) with the ratification of the RKUHP. In this phase, irrelevant tweets are also deleted.

#### C. Data Preprocessing

In the preprocessing of tweet data, a series of operations are performed so that machine learning algorithms can read the tweet's standards and patterns in Table II. The method is as follows:

- 1) Case Folding turns all capital characters in tweets into lowercase letters.
- 2) Remove Punctuation and eliminates punctuation, URL links, numbers, hashtags, and emoticons from tweets.
- 3) Tokenization is the process of breaking sentences into separate words.
- 4) Stop Word is the process of removing words that don't add any meaning.
- 5) Normalization is the process of uniforming words with the same meaning but different spellings.
- 6) Stemming is changing words that have affixes into essential words.

TABLE II. PREPROCESSING PROCESS

Process	Tweet
User Tweet	Dukung pengesahan RKUHP untuk supremasi hukum https://t.co/MLmR6BHBIS #DukungPengesahanKUHP
Case folding	dukung pengesahan rkuhp untuk supremasi hukum https://t.co/mlmr6bhbis #dukungpengesahankuhp
Remove Punctuation	dukung pengesahan rkuhp untuk supremasi hukum
Tokenizing	['dukung', 'pengesahan', 'rkuhp', 'untuk', 'supremasi', 'hukum']
Stop Word	['dukung', 'pengesahan', 'rkuhp', 'supremasi', 'hukum']
Normalization	['dukung', 'pengesahan', 'rkuhp', 'supremasi', 'hukum']
Stemming	['dukung', 'kesah', 'rkuhp', 'supremasi', 'hukum']

#### D. N-Gram

In this phase, word separation is carried out; we combine the unigram, the bigram, and the trigram forms of the n-gram. Words are created using unigrams (one word), bigrams (two words), and trigrams (three words). Tweets that have at most three words will be deleted.

#### E. Term Weighting

The next step was word feature extraction using the term frequency-inverse document frequency (TF-IDF). The word weight in a given document is typically calculated using the TF-IDF technique. The term frequency describes the often appearing; that often appears in a manuscript (TF). Frequently occurring terms will obstruct the search for uncommon words. The inverse document frequency (IDF), which lessens the weight of often-appearing words, can gauge how significant a word's meaning is in a document [31].

#### F. Classification Modelling

In this step, classification modeling is applied to the test data using three machine learning algorithms: Nave Bayes, SVM, and CART. Modeling is done separately to produce accurate results. Each algorithm tests the words formed in the ngram process, and the term weighting process has been carried out. This classifier uses the sklearn library in Python. This study used 80% training data and 20% testing data. This is so that machine learning algorithms can perform better, according to research by Pham et al. in Nguyen et al. research, when training data is raised from 30% to 80%. However, when it is increased from 80% to 90%, the opposite occurs[32] [33].

G. Classification Modelling Evaluation

In this phase, the performance of each machine learning algorithm in the previous step will be evaluated. Evaluation is conducted using a confusion matrix by looking at the value of the accuracy of each algorithm. Accuracy, precision, and recall are the evaluative test parameters whose computations are derived from the confusion matrix table [13].

H. Sentiment Result and Discussion

This is the final stage in producing sentiment words for the word cloud. Which, according to the N-gram phase, consists of one word, two words, and three terms and is derived from the sentiment with the maximum accuracy. Then, the discussion will be made in light of these findings.

IV. RESULT AND DISCUSSION

A. Results of Classification Modeling Evaluation

The number of tweets extracted using the snsrape library is 17,107. After cleaning the same tweets, the number of tweets increases to 10,763. Then, label each tweet manually. Then, preprocessing process the tweet and generate it again to yield 9,079 tweets. The tweet then executed the classification algorithms and the n-gram combination method. After preprocessing, the dataset is split into training and test sets. 80% of the dataset is used for training, and the remaining 20% is used for testing. The dataset's features are produced using an n-gram mix of unigrams, bigrams, and trigrams. The created words will then be weighted using term analysis. Different data are made when n-grams and term weighting are combined. The results are presented in Tables III, IV, and V.

Using the confusion metrics, we have calculated the performance of each algorithm here. The confusion matrix, which measures the classification overlap, is an effective tool for performance evaluation. The multi-label classification task must establish the confusion matrix because each instance may be assigned to multiple classes [34]. The performance evaluation of the multi-label classifier is based on computing performance averages, including precision, recall, and F1-score [34]. Precision measures how accurate a class's predictions are relative to all the predictions included in the course. Recall is the percentage of a class's total number of categorized facts that can be predicted accurately. The f1 score calculation was then utilized to mix the precision and recall [35] [12].

For each n-gram combination used, precision, recall, and f1-scores for the CART algorithm are displayed in table III. The findings of CART do not differ much when unigrams, bigrams, or trigrams are used. In the bigram findings, for example, the precision value is 0.73 for negative and 0.75 for positive, and the recall value is 0.70 for negative and 0.75 for positive. The f1 values for positive and negative are then 0.72 and 0.74, respectively. As shown in Fig. 2, out of the 852 negatively judged tweets, 624 were true negatives (TN), and 228 were false negatives. In contrast, out of 964 positive tweets, 263 were false positives (FP), and 701 were true positives (TP).

TABLE III. PRECISION, RECALL, AND F1-SCORE CART

	N-Gram	Precision	Recall	F1-score
Negative	Unigram	0.74	0.70	0.72
	Bigram	0.73	0.70	0.72
	Trigram	<b>0.75</b>	0.69	0.72
Positive	Unigram	0.72	0.76	0.74
	Bigram	0.73	0.75	0.74
	Trigram	<b>0.71</b>	<b>0.77</b>	0.74

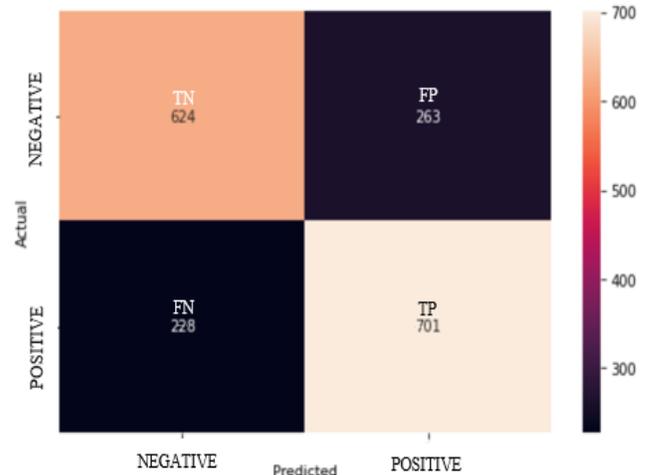


Fig. 2. The confusion matrix bigram CART

Table IV shows the SVM algorithm's precision, recall, and f1-score for each n-gram combination. The unigram test had the best average outcomes, with precision values of 0.81 for negative and 0.80 for positive groups and recalled values of 0.76 for negative and 0.82 for positive. The f1-score is 0.78 for the negative and 0.81 for the positive. As shown in **Error! Reference source not found.**, 711 of the 877 tweets that received a negative evaluation were true negatives, and 166 were false negatives. Comparatively, out of 939 positive tweets, 186 were false positives, and 753 were true positives.

TABLE IV. PRECISION, RECALL, AND F1-SCORE SVM

	N-Gram	precision	recall	f1-score
Negative	Unigram	<b>0.81</b>	0.79	<b>0.80</b>
	Bigram	0.79	<b>0.78</b>	0.79
	Trigram	0.81	0.76	0.78
Positive	Unigram	<b>0.80</b>	<b>0.82</b>	<b>0.81</b>
	Bigram	0.79	0.81	0.80
	Trigram	0.78	0.82	0.80



Fig. 3. The confusion matrix unigram SVM

The Naïve Bayes algorithm is presented in Table V for each possible combination of n-grams. The trigram test had the most outstanding results overall, with accuracy scores of 0.78 for negative responses and 0.79 for positive ones, recall scores of 0.79 for negative answers and 0.77 for positive reactions, and an f1-score of 0.78 for negative and positive responses. As shown in **Error! Reference source not found.**, of the 926 tweets that received a negative rating, 718 were true negatives, and the remaining 208 were false negatives. The 890 positive tweets, in contrast, contained 186 false positives and 704 true positives.

TABLE V. PRECISION, RECALL AND F1-SCORE NAÏVE BAYES

	N-Gram	precision	recall	f1-score
Negative	Unigram	0.75	0.77	0.76
	Bigram	0.75	0.79	0.77
	Trigram	0.78	0.79	0.78
Positive	Unigram	0.77	0.75	0.76
	Bigram	0.79	0.75	0.77
	Trigram	0.79	0.77	0.78



Fig. 4. The confusion matrix trigram naïve bayes

Calculating the accuracy of each method is another function of the confusion matrix, which can be seen in Table VI. It has been demonstrated that the SVM constructed using the unigram has the maximum accuracy, equal to 0.81. In addition, bigram and trigram SVM continues to have the highest accuracy compared to other algorithms, with respective values of 0.79 and 0.78. When utilizing trigram combinations, Naïve Bayes on 0.78 achieves a higher level of accuracy. CART has the same accuracy in all ngram combinations.

TABLE VI. ACCURACY CART, SVM AND NAÏVE BAYES

Algorithm	NGRAM Accuracy		
	Unigram	Bigram	Trigram
CART	0.73	0.73	0.73
SVM	0.81	0.79	0.79
Naïve Bayes	0.76	0.77	0.78

The analysis results in Table VI are consistent with research in [13] and [15] that shows that SVM has higher accuracy when compared to Naïve Bayes and CART. The SVM study achieved an accuracy of 83.34% and a Naïve Bayes of 75%. According to research [15], the amount of data used by the random forest and CART algorithms determines the soundness of multiple decision trees, the complexity of the trees, and thus the algorithm's accuracy. This explains why CART has the same accuracy because it has the same number of data sets.

In line with the results of this investigation, a study in [16] discovered that SVM with the unigram combination had the highest accuracy when compared to the other ngram combinations. This is likely due to the ease with which SVM can map words weighted with TF-IDF rather than utilizing multiple words to infer sentiment. This compares favorably with naïve Bayes, where the more word combinations in ngrams, the higher the accuracy. Many ngram combinations raise the level of accuracy in Naïve Bayes. Therefore, based on this, it was found that the combination would affect the accuracy of each algorithm. SVM is preferred over algorithms, Naïve Bayes, and CART because of its high accuracy. For Naïve Bayes, a higher gram would be preferable. Because CART is affected by a large amount of data, vast amounts of data will affect its accuracy.

B. Content Sentiment Analysis

Nine thousand seventy-nine tweets were included in the data obtained after being processed using Python and Microsoft Excel programming languages. Several duplicate and irrelevant tweets have been removed from the message. The result is that 51% of tweets, or 4.623 of them, favor the ratification of the RKUHP, while 49% of tweets, or as many as 4.455 of them, are in opposition to it, as can say be seen in **Error! Reference source not found.**

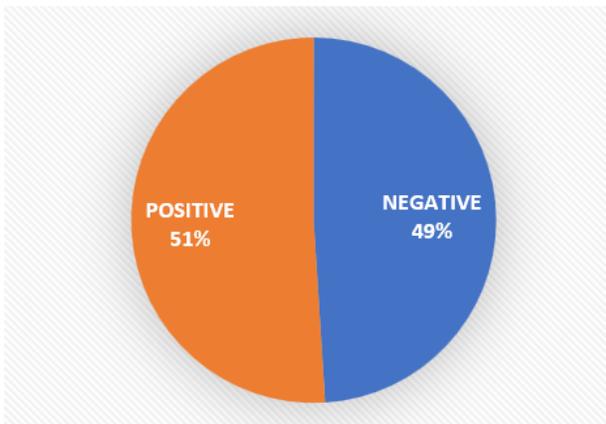


Fig. 5. Result of the sentiment of the ratification of RKUHP

The word cloud for a negative sentiment is displayed in **Error! Reference source not found.** Negative sentiment is associated with a wide variety of topics and concepts, including 'tolak', 'kontroversi', 'kritik', 'hina', 'koruptor', 'penting kuasa'. The most common words are 'sah', 'tolak,' 'rakyat,' and 'negara' in that order. The RKUHP received a negative response because it was thought to contain several articles that could be construed as contentious. Based on word cloud sentiments such as the words 'kritik,' 'hina' and 'demokrasi', some articles are considered to silence criticism, specifically regarding insulting the president. Then the words 'koruptor' and 'korupsi' articles regarding corruption, with a minimum reduction in prison for corruption. Then there's the phrase 'penting kuasa' and 'rakyat', because some people believe that a lot of the new RKUHP articles were written more for the authorities' interests than for the people's interests themselves.



Fig. 6. Word cloud of negative sentiment

Besides having negative sentiments, there are also words representing positive sentiments in the word cloud. Word like 'sah', 'baru', 'hukum pidana', 'baru' dan 'tujuan' support the approval of the ratification of this RKUHP. The RKUHP is significant because it strengthens Indonesia's current criminal code. In Indonesia, criminal law that has undergone patchwork is no longer regarded as complying with legal criteria. The positive word Cloud is shown in **Error! Reference source not found.**



Fig. 7. Word cloud of positive sentiment

We may deduce what words are at the center of people's conversations based on the outcomes of positive and negative sentiments in **Error! Reference source not found.** and **Error! Reference source not found.** The words that arise may serve as a first reflection for the organization of positive and negative things that are the community's response. This can be used as a resource for organizations to improve their understanding of the policies they issue. This is in line with the findings of a study [8] on how the Mexican government uses Twitter to connect with the people.

As a result, the outcomes of these attitudes can be employed by the government as a foundation for socialization. Because there is still the problem of the pessimism of RKUHP, there are still drawbacks to the ratification of the RKUHP, which is still relatively high and reaches 49% of the population. To find a solution to this problem, the government needs to engage in more social activities and listen to people's perspectives on matters that are regarded as contentious. This is done to ensure that both the adoption of the RKUHP in 2025 and its passage into the Criminal Code happen smoothly. Words that elicit negative responses might be utilized as the primary focus of socialization. This will help in mitigating the public's adverse reaction. Mitigation of this rejection will be better if there is a grouping of tweets based on topic, as in Research [36]. In this study, we used BERTopic to classify the tweets. BERTopic will help categorize tweets and make it easier for the government to bring up specific RKUHP-related topics. Throughout the processing, it was discovered that the steaming method had limitations since particular terms, such as "pengesahan", were mistakenly converted into the root word "kesah". Hence the potential limits of the world cloud's word interpretation.

## V. CONCLUSION

Examining Twitter sentiment, this study identifies responses to the ratification of the RKUHP. The RKUHP ratification drew 51% positive and 49% negative comments on Twitter. This demonstrates that, even though the positive is superior, the value is just 2%. According to the negative comments, the problem of controversial articles is related to the article about insulting the president, the post about cutting punishments for corrupt officials, and the piece about not representing the people. This must be the emphasis of the government's efforts to socialize the RKUHP.

The evaluation of the three tested algorithms—CART, SVM, and Naïve Bayes—found that SVM had the highest accuracy and was the most reliable even when the n-gram combination was used. SVM produces an accuracy value of 0.81 on the unigram, 0.79 on the bigram, and 0.79 on the trigram.

This research is limited to grouping tweets that have yet to be grouped into specific topics and imperfections in the steaming process. It is hoped that future research can categorize recent tweets based on grouping relevant issues related to the RKUHP so that they are not only the results of grouping terms from the Word Cloud. It can also add more data which makes the topic even better. In addition, it can improve the algorithm steaming process to make it better.

#### ACKNOWLEDGMENT

Thanks to the Indonesian Ministry of Communication and Information for supporting, assisting, and funding this research. As well as support from the Faculty of Computer Science, University of Indonesia.

#### REFERENCES

- [1] Humas dan Protokol BPHN, "RUU KUHP Disahkan menjadi Undang-undang," 2022. <https://bphn.go.id/publikasi/berita/202212061210189/ruu-kuhp-disahkan-menjadi-undang-undang> (accessed Dec. 16, 2022).
- [2] Y. Y. WEDHA and E. NURCAHYO, "Criminal Law Reform Toward Deprivation of Property Resulting from Corruption Criminal Acts," *PRIZREN Soc. Sci. J.*, vol. 5, no. 1, pp. 97–103, Apr. 2021, doi: 10.32936/pssj.v5i1.207.
- [3] M. F. Butar-butur, "Exemptions from Liability in Indonesian Criminal Law Reform," *Ann. R.S.C.B.*, vol. 25, no. 5, pp. 5528–5533, 2021.
- [4] "UU KUHP Telah Akomodir Seluruh Aspirasi Masyarakat Indonesia," 2022. <https://www.dpr.go.id/berita/detail/id/42227/t/uu+kuhp+telah+akomodir+seluruh+aspirasi+masyarakat+indonesia> (accessed Dec. 16, 2022).
- [5] S. Strangio, "Indonesia Set to Pass Controversial New Criminal Code This Month," *The Diplomat*, Dec. 2022. [Online]. Available: <https://www.proquest.com/magazines/indonesia-set-pass-controversial-new-criminal/docview/2745669907/se-2>.
- [6] A. Alsaedi and M. Z. Khan, "A study on sentiment analysis techniques of Twitter data," *Int. J. Adv. Comput. Sci. Appl.*, vol. 10, no. 2, pp. 361–374, 2019, doi: 10.14569/ijacsa.2019.0100248.
- [7] J. K. Harris, S. Moreland-Russell, B. Choucair, R. Mansour, M. Staub, and K. Simmons, "Tweeting for and against public health policy: Response to the Chicago Department of Public Health's electronic cigarette Twitter campaign," *J. Med. Internet Res.*, vol. 16, no. 10, p. e238, 2014, doi: 10.2196/jmir.3622.
- [8] R. B. Hubert, E. Estevez, A. Maguitman, and T. Janowski, "Analyzing and Visualizing Government-Citizen Interactions on Twitter to Support Public Policy-making," *Digit. Gov. Res. Pract.*, vol. 1, no. 2, pp. 1–20, 2020, doi: 10.1145/3360001.
- [9] M. A. Kausar, A. Soosaimanickam, and M. Nasar, "Public Sentiment Analysis on Twitter Data during COVID-19 Outbreak," *Int. J. Adv. Comput. Sci. Appl.*, vol. 12, no. 2, pp. 415–422, 2021, doi: 10.14569/ijacsa.2021.0120252.
- [10] S. H. A. Samsudin, N. M. Sabri, N. Isa, and U. F. M. Bahrain, "Sentiment Analysis on Acceptance of New Normal in COVID-19 Pandemic using Naïve Bayes Algorithm," *Int. J. Adv. Comput. Sci. Appl.*, vol. 13, no. 9, pp. 581–588, 2022, doi: 10.14569/ijacsa.2022.0130968.
- [11] M. Ahmad, S. Aftab, M. S. Bashir, N. Hameed, I. Ali, and Z. Nawaz, "SVM optimization for sentiment analysis," *Int. J. Adv. Comput. Sci. Appl.*, vol. 9, no. 4, pp. 393–398, 2018, doi: 10.14569/ijacsa.2018.090455.
- [12] A. N. Muhammad, S. Bukhori, and P. Pandunata, "Sentiment Analysis of Positive and Negative of YouTube Comments Using Naïve Bayes-Support Vector Machine (NBSVM) Classifier," *Proc. - 2019 Int. Conf. Comput. Sci. Inf. Technol. Electr. Eng. ICOMITEE 2019*, pp. 199–205, 2019, doi: 10.1109/ICOMITEE.2019.8920923.
- [13] R. Kusumawati, A. D'Arofah, and P. A. Pramana, "Comparison Performance of Naive Bayes Classifier and Support Vector Machine Algorithm for Twitter's Classification of Tokopedia Services," *J. Phys. Conf. Ser.*, vol. 1320, no. 1, 2019, doi: 10.1088/1742-6596/1320/1/012016.
- [14] I. C. Sari and Y. Ruldeviyani, "Sentiment Analysis of the Covid-19 Virus Infection in Indonesian Public Transportation on Twitter Data: A Case Study of Commuter Line Passengers," *2020 Int. Work. Big Data Inf. Secur. IWBIS 2020*, pp. 23–28, 2020, doi: 10.1109/IWBIS50925.2020.9255531.
- [15] N. Leelawat, S. Jariyapongpaiboon, A. Promjun, and S. Boonyarak, "Heliyon Twitter data sentiment analysis of tourism in Thailand during the COVID-19 pandemic using machine learning," *Heliyon*, vol. 8, no. September, p. e10894, 2022, doi: 10.1016/j.heliyon.2022.e10894.
- [16] T. Sinnasamy and N. N. A. Sjaif, "Sentiment Analysis using Term based Method for Customers' Reviews in Amazon Product," *Int. J. Adv. Comput. Sci. Appl.*, vol. 13, no. 7, pp. 685–691, 2022, doi: 10.14569/ijacsa.2022.0130780.
- [17] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?," in *Proceedings of the ACL-02 conference on Empirical methods in natural language processing - EMNLP '02*, 2002, vol. 10, pp. 79–86. doi: 10.3115/1118693.1118704.
- [18] S. B. Kotsiantis and D. Kanellopoulos, "Data preprocessing for supervised learning," *Int. J. ...*, vol. 1, no. 2, pp. 1–7, 2006, doi: 10.1080/02331931003692557.
- [19] P. Yerpude and V. Gudur, "Predictive Modelling of Crime Dataset Using Data Mining," *Int. J. Data Min. Knowl. Manag. Process*, vol. 7, no. 4, pp. 43–58, 2017, doi: 10.5121/ijdkp.2017.7404.
- [20] O. Oriola and E. Kotze, "Evaluating Machine Learning Techniques for Detecting Offensive and Hate Speech in South African Tweets," *IEEE Access*, vol. 8, pp. 21496–21509, 2020, doi: 10.1109/ACCESS.2020.2968173.
- [21] K. S. Nugroho et al., "Detecting Emotion in Indonesian Tweets: A Term-Weighting Scheme Study," *Int. J. Adv. Comput. Sci. Appl.*, vol. 13, no. 1, pp. 61–70, 2022, doi: 10.20473/jisebi.8.1.61-70.
- [22] K. SPARCK JONES, "A STATISTICAL INTERPRETATION OF TERM SPECIFICITY AND ITS APPLICATION IN RETRIEVAL," *J. Doc.*, vol. 28, no. 1, pp. 11–21, Jan. 1972, doi: 10.1108/eb026526.
- [23] Kowsari, Jafari Meimandi, Heidarysafa, Mendu, Barnes, and Brown, "Text Classification Algorithms: A Survey," *Information*, vol. 10, no. 4, p. 150, Apr. 2019, doi: 10.3390/info10040150.
- [24] F. A. Bachtiar, W. Paulina, and A. N. Rusydi, "Text Mining for Aspect Based Sentiment Analysis on Customer Review : a Case Study in the Hotel Industry," *5th Int. Work. Innov. Inf. Commun. Sci. Technol.*, no. March, 2020.
- [25] F. Razaque et al., "Using naïve bayes algorithm to students' bachelor academic performances analysis," in *2017 4th IEEE International Conference on Engineering Technologies and Applied Sciences (ICETAS)*, Nov. 2017, pp. 1–5. doi: 10.1109/ICETAS.2017.8277884.
- [26] J. Han, M. Kamber, and J. Pei, "Data Mining: Concepts and Techniques," *Third.*, Boston: Elsevier, 2012, pp. 408–415. doi: 10.1016/B978-0-12-381479-1.00027-7.
- [27] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*. Routledge, 1984. doi: 10.1201/9781315139470.
- [28] B. Choubin, G. Zehtabian, A. Azareh, E. Rafiei-Sardooi, F. Sajedi-Hosseini, and Ö. Kişi, "Precipitation forecasting using classification and regression trees (CART) model: a comparative study of different approaches," *Environ. Earth Sci.*, vol. 77, no. 8, pp. 1–13, 2018, doi: 10.1007/s12665-018-7498-z.
- [29] R. Singh, T. Wagener, R. Crane, M. E. Mann, and L. Ning, "A vulnerability driven approach to identify adverse climate and land use change combinations for critical hydrologic indicator thresholds: Application to a watershed in Pennsylvania, USA," *Water Resour. Res.*, vol. 50, no. 4, pp. 3409–3427, Apr. 2014, doi: 10.1002/2013WR014988.

- [30] B. Choubin, H. Darabi, O. Rahmati, F. Sajedi-Hosseini, and B. Kløve, "River suspended sediment modelling using the CART model: A comparative study of machine learning techniques," *Sci. Total Environ.*, vol. 615, pp. 272–281, Feb. 2018, doi: 10.1016/j.scitotenv.2017.09.293.
- [31] Raksaka Indra Alhaqq, I Made Kurniawan Putra, and Yova Ruldeviyani, "Analisis Sentimen terhadap Penggunaan Aplikasi MySAPK BKN di Google Play Store," *J. Nas. Tek. Elektro dan Teknol. Inf.*, vol. 11, no. 2, pp. 105–113, 2022, doi: 10.22146/jnteti.v11i2.3528.
- [32] B. T. Pham et al., "A Novel Hybrid Soft Computing Model Using Random Forest and Particle Swarm Optimization for Estimation of Undrained Shear Strength of Soil," *Sustainability*, vol. 12, no. 6, p. 2218, Mar. 2020, doi: 10.3390/su12062218.
- [33] Q. H. Nguyen et al., "Influence of Data Splitting on Performance of Machine Learning Models in Prediction of Shear Strength of Soil," *Math. Probl. Eng.*, vol. 2021, pp. 1–15, Feb. 2021, doi: 10.1155/2021/4832864.
- [34] M. Heydarian, T. E. Doyle, and R. Samavi, "MLCM: Multi-Label Confusion Matrix," *IEEE Access*, vol. 10, pp. 19083–19095, 2022, doi: 10.1109/ACCESS.2022.3151048.
- [35] D. H. Wahid and A. SN, "Peringkasan Sentimen Esktraktif di Twitter Menggunakan Hybrid TF-IDF dan Cosine Similarity," *IJCCS (Indonesian J. Comput. Cybern. Syst.)*, vol. 10, no. 2, p. 207, Jul. 2016, doi: 10.22146/ijccs.16625.
- [36] F. Alhaj, A. Al-Haj, A. Sharieh, and R. Jabri, "Improving Arabic Cognitive Distortion Classification in Twitter using BERTopic," *Int. J. Adv. Comput. Sci. Appl.*, vol. 13, no. 1, pp. 854–860, 2022, doi: 10.14569/IJACSA.2022.0130199.

# Evaluation of QoS over IEEE 802.11 Wireless Network in the Implementation of Internet Protocols Mobility Supporting

Narimane Elhilali<sup>1</sup>, Mostapha Badri<sup>2</sup>, Mouncef Filali Bouami<sup>3</sup>

Applied Mathematics and Information Systems Laboratory-Multidisciplinary Faculty of Nador,  
University Mohammed Premier Oujda, Morocco<sup>1,2,3</sup>

**Abstract**—Now-a-days, the internet is an essential part of our digital lives. With the growing number of users, the ultimate goal is to enable all users to stay connected to the internet at anytime and anywhere, regardless of their mobility. Any delay or jitter in the system can cause a deterioration in the performance of multimedia services, such as video streaming, or cause websites to partially load. The current Internet Protocol version 4 (IPv4) cannot handle all the IP addressing requirements, while the next generation Internet Protocol version 6 (IPv6) has been developed to solve some of these problems by improving the quality of service and providing many other features. The primary contribution of this paper is to investigate the evaluation of Quality of Service (QoS) functionality, including end-to-end delay, throughput, jitter, and packet loss, in WLAN mobility environments for MIPv4 to MIPv6 using the OMNeT++ simulator.

**Keywords**—QoS; MIPv4; MIPv6; handover; mobility; priority

## I. INTRODUCTION

In recent years, communication networks have become the main engine of the world, especially mobile communications, which have spread everywhere. Wireless networks that provide access to the internet are the solution for users who need to move around while exchanging multimedia services such as video, voice, or data. The IEEE 802.11 standard [1], which includes the medium access control (MAC) and physical layer (PHY), has been improved since its initial approval by the IEEE in 1997. The improvements are defined as amendments to the initial standard, such as IEEE 802.11e and IEEE 802.11r, which mainly concern the MAC layer. IEEE 802.11e added QoS mechanics, while IEEE 802.11r improved mobility between the cells of wireless networks (handover) [2] and allowed nodes to switch more quickly between access points (APs).

The Mobile Internet Protocol (MIP) [3] is an IETF communication protocol that allows users to stay connected when moving from one cellular network to another without interrupting the connection. The two versions of MIP, Mobile IPv4 (RFC 3344) and Mobile IPv6 (RFC 3775) [4], are designed to facilitate node mobility and maintain connections when changing locations. MIPv4 is a popular mobile internet protocol based on the IPv4 protocol, which has the responsibility of traffic routing on the internet. However, it cannot handle the IP addressing margin for all users and also has some limitations in terms of quality of service, MIPv6 has

been developed to provide mobility support for IPv6, and improves some of this limitations. In order to ensure node mobility, applications and flows, quality of service is particularly important for traffic, and it can give priority to different flow services (such as video, voice, data, etc.) based on a certain level of performance. QoS metrics [5] measure performance in four different aspects: throughput, delay, jitter, and packet loss.

The main objective of this article is to assess the impact of internet protocol mobility support, from Mobile IPv4 to Mobile IPv6, on QoS performance in mobility environments based on 802.11g wireless networks using the OMNeT++ simulator [6]. The QoS functionalities are analyzed and compared, the packets are classified into four traffic classes (background, best effort, video, and voice) with different priority levels of the EDCA mechanism according to the nominal bit rate.

The rest of this paper is organized as follows. Section (2) describes related works on QoS performance on mobile internet protocols. Section (3) introduces QoS in IEEE802.11. Section (4) provides the different QoS Parameters. The simulation results include Simulation Environment, Simulation Scenario, results, analysis, and simulation comparison are talked in section (5), finally we conclude in section (6) and future work in section (7).

## II. RELATED WORKS

Several studies have been conducted to evaluate the quality of service (QoS) performance of different network protocols, including Internet Protocol version 6 (IPv6) and Mobile Internet Protocol version 6 (MIPv6). In this paragraph, we discuss three studies that evaluate the QoS performance of video and audio applications in IPv6 and MIPv6 using simulation tools such as OPNET. The studies analyze various QoS metrics such as delay variation, end-to-end delay, packet loss, and handover latency to investigate the effectiveness of different network standards and protocols. The results of these studies provide valuable insights into the strengths and weaknesses of different protocols in terms of QoS performance.

E.S. Ikeremo and M.C. Kelly T. Pepple [7] discuss the evaluation of QoS performance metrics of video streaming in IPv6, including delay variation, end-to-end delay, and packet loss. The simulation analyzes the effect of frame rates, type of

service, and bandwidth parameters on the QoS metrics using the OPNET environment.

In [8], the authors study the evaluation of QoS in Mobile Internet Protocol v6 using IEEE 802.11e and IEEE 802.11b standards by OPNET simulator. The paper based on the Route Optimization to investigate the QoS metrics such as packet delay variation, HA binding delay, and latency in the MIPv6 handover with video conference applications in real time. The results indicate that the IEEE 802.11e amendment is more effective than IEEE 802.11b during the handover process.

Zakari et al. [9] present a comparative performance study of IPv4 and IPv6 protocols based on the results of QoS metrics of video and audio applications. The study shows that IPv6 performs better than IPv4 in both scenarios.

### III. QoS IN IEEE 802.11

IEEE 802.11e is an amendment to the IEEE 802.11 standard that was approved in 2005. It introduces Quality of Service (QoS) enhancements to the Medium Access Control (MAC) protocol sub-layer of the data link layer of the OSI model. 802.11e improves WLANs by enabling the transport of voice and video with QoS. The packets can belong to different traffic classes that have different transmission priorities. Packets with high priority are more likely to be transmitted before lower priority packets, which reduces delay and jitter for responsive applications.

When QoS is implemented in 802.11, the MAC uses different technical functions [10]. One of these techniques is Enhanced Distributed Channel Access (EDCA), which provides different processing for different classes of packets and channel capacities. EDCA is part of the Hybrid Coordination Function (HCF) and defines four categories of channel access (or priorities), as shown in the table below, from the lowest to the highest priority categories. The Table I represent the different Access categories priority for EDCA function:

TABLE I. ACCESS CATEGORIES PRIORITY IN EDCA FUNCTION

Level	Priority	Access Category
Lowest	1	Background
	2	
to	0	Best Effort
	3	
	4	Video
	5	
Highest	6	Voice
	7	

### IV. QoS PARAMETERS

IPv6 is the most recent version of Internet Protocol [11], this new IP address was involved to fulfill the need for more Internet addresses, and to treat some requirements and limits of IPv4. The quality of service is the important requirement in this new version of the protocol, it has been supported and improved. The evaluation of QoS performance can be measured based on different parameters [12], such as:

#### A. Network Throughput

Throughput is the maximum transmission capacity of a volume of data between two points on a communication line in a given time. QoS optimizes the network by managing network bandwidth and prioritizing applications according to the resources they need. The mathematical formula for throughput is:

$$\text{Throughput} = \frac{\text{Packets Received of data}}{\text{delivery time}} \quad (1)$$

#### B. End-to-End Delay

End-to-end delay is the time it takes for a packet to travel from the source to the destination. It should ideally be as close to zero as possible. It can also be defined as the time difference between the instance of sending and receiving of the packet between two nodes. The mathematical formula for end-to-end delay is:

$$\text{Delay} = \frac{\sum \text{link packet delays}}{\sum \text{packets received}} \quad (2)$$

Where the sum of link packets delays is:

$$d_{\text{transmission}} + d_{\text{propagation}} + d_{\text{queueing}} + d_{\text{processing}} \quad (3)$$

$$d_{\text{trans}} = \frac{\text{packet length}}{\text{transmission rate}} \quad (4)$$

$$d_{\text{prop}} = \frac{\text{distance btw nodes}}{\text{propagation rate}} \quad (5)$$

$d_{\text{queue}}$  depends on congestion and  $d_{\text{proces}}$  is a few microseconds.

#### C. Network Jitters

Jitter, also known as Packet Delay Variation (PDV), is a phenomenon that occurs due to network congestion or queuing, or when data packets are delayed or lost. If jitter is too high, it can lead to a deterioration in the quality of voice or audio communication. In OMNeT++, packet jitter is measured as the difference between the packet delays of successive packets, which is called the Instantaneous Packet Delay Variation. The mathematical formula for jitter is:

$$\text{Jitter} = \frac{\sum_i^n D_i}{\sum \text{packets received}} \quad (6)$$

Where, the sum of delay variation D is:

$$(D_2 - D_1) + \dots + (D_n - D_{n-1}) \quad (7)$$

D. Packets Loss

Packet loss is defined as the number of data packets that are dropped between two nodes during network traffic. The mathematical formula for packet loss is:

$$\text{PacketLoss} = \frac{\text{Packets Sent} - \text{Packets Dropped}}{\text{Packets Sent}} \times 100\% \quad (8)$$

V. SIMULATION RESULTS

This part, describe the simulation paradigm that was studied in OMNeT++, the simulation scenarios that were chosen, and the results obtained from the experiments.

A. Simulation Environment

In this section, we describe the environment for our simulation experiments, which was implemented using the OMNeT++ 5.6.4 simulator. Our simulation represents a wireless network based on the IEEE 802.11e, IEEE 802.11g, and IEEE 802.11r standards. It includes wireless hosts moving throughout the network area, separated to trigger handover and communicate via access points. The wireless host equipment used in the simulation is a compatible node with support for the IPv6 protocol, as well as handover mechanisms and the Mobile IPv6 protocol. The access points used support multiple wireless radios and multiple Ethernet ports. Table II outlines the simulation parameters, and Table III lists the applications used during the simulation.

TABLE II. NETWORK SIMULATION PARAMETERS

Parameters	Value
Network Simulator	OMNeT++ (V 5.6.2)
Framework	INET
Simulation Area	600x400
Simulation Time	10 s
Channel	Wireless (IEEE802.11)
Standard	IEEE802.11 e/g/r
Speed of node	10 mps
Mobility Model	Linear Mobility
Application layer	TCP, UDP
Network interface model	PHY/WIFI's MAC
Internet Protocol	IPv4, IPv6
Frequency	2.412 GHz
Bandwidth	20 MHz
Data Rate	54Mbit/s
AP Beacon Interval	100 ms
Performance streams	End-to-end delay, packet delay variation, throughput, Packet Loss

TABLE III. NETWORK SIMULATION APPLICATION PARAMETERS

Access Category	Packet Length	Packets Access Priority	Nominal Bitrate
Background	900 B	1	24 Mbps
Best Effort	900 B	0	28 Mbps
Video	600 B	5	5 Mbps
Audio	125 B	6	100 kbps

B. Simulation Scenario

The simulation in Fig. 1 demonstrates handover between two access points (APs) in an 802.11g wireless LAN. A wireless node (sender) moves linearly through the network at a speed of 10 m/s, while the wireless node (receiver) remains stationary. Both nodes are configured to use a PHY rate of 54 Mbit/s. The two access points are separated by a distance of 400 meters. When the host moves within the network area, it uses an active scanning method to attach to the nearest AP, choosing the one with the highest signal strength before exchanging data. Two simulation scenarios were created, one with the wireless node implemented with IPv4 protocol mobility support, and the other with the node implemented to support mobility in IPv6. In both simulation scenarios, the source node sends UDP data to the destination node in wireless mode via four UDP streams, each corresponding to a different access category (background, best effort, video, and audio). QoS functionality is enabled, and parameters such as end-to-end delay, jitter, throughput, and packet loss are measured and analyzed to examine how mobile protocols affect the performance of each other for different access categories during horizontal handover. Packets with the highest priority must have lower delay times and higher throughput. The figures below show the network design implemented in the OMNeT++ simulator and the flow chart of the simulation scenario.

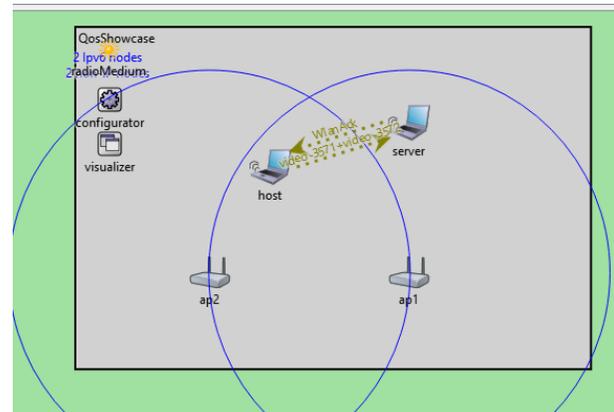


Fig. 1. Network topology design in OMNeT++

The overall flow chart of our proposed scenario is shown as Fig. 2.

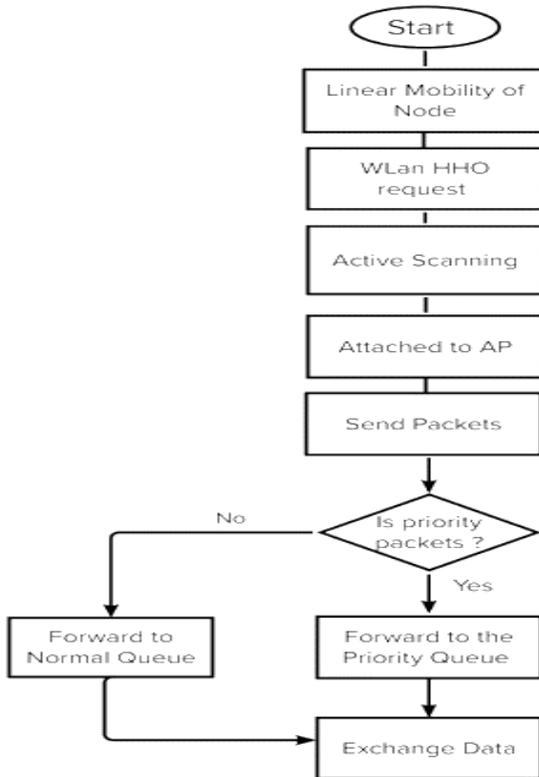


Fig. 2. Flow chart of simulation scenario proposed

C. Results and Analysis

1) Scenario 1: It was implemented with IPv4 mobility support to evaluate the QoS functionality between the two nodes during handover of the two access points in a wireless network.

In Fig. 3, we observed that the throughput values matched their nominal bitrate for high priority traffic (video and audio), 5Mbps for video and 100kbps for audio, in contrast to the lower priority traffic (background and best effort), which had lower throughput values. This explains the instability in the graphs.

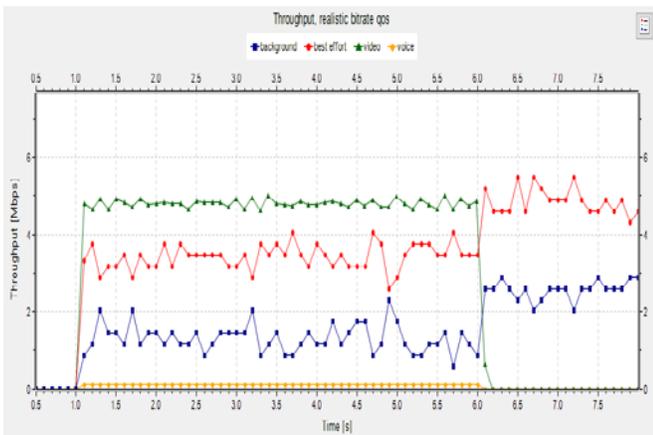


Fig. 3. Throughput variations in scenario 1

The Fig. 4 presents the jitter values for the video and audio access categories remained relatively low (especially for video) compared to the more dispersed scatter points observed for the background and best effort categories, which can reach up to 0.06s. Jitter values began to decrease when the video and audio traffic stopped.

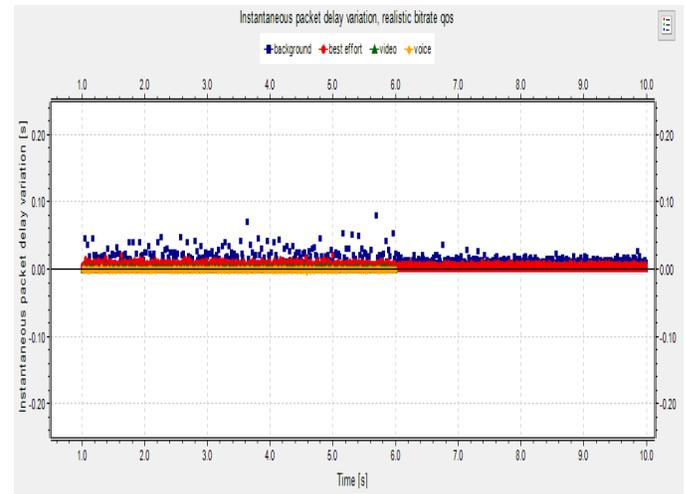


Fig. 4. Jitter variations in scenario 1

This scatter plot in Fig. 5 analyzes the delay of each of the four access categories studied. The video and audio categories were more likely to be sent first, which explains the lower values of packet delay observed in these categories. The best effort category was prioritized over background because its packets were sent periodically, which takes more time.

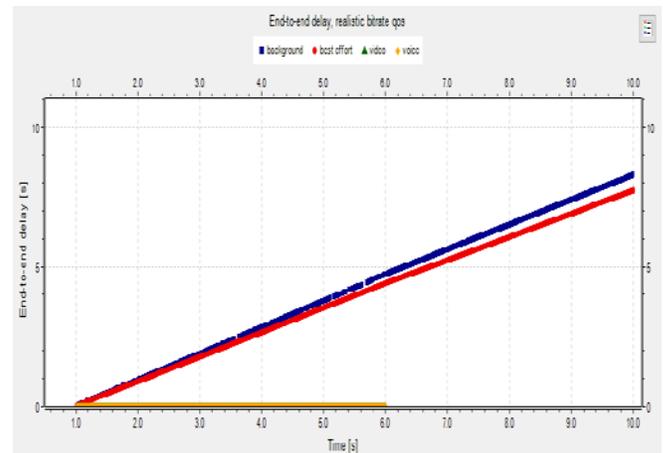


Fig. 5. End-to-end delay variations in scenario 1

In Scenario 1, the QoS parameters were measured for various access categories. The results were summarized in Table IV that provides values for parameters such as end-to-end delay, jitter, throughput, and packet loss, and allows us to compare the performance of different categories in terms of QoS. Overall, the results suggest that high-priority traffic, such as video and audio, had a lower jitter and delay compared to low-priority traffic, such as background and best effort.

TABLE IV. SIMULATION RESULTS OF QoS IN IPV4 MOBILITY SUPPORT

Access Category	Throughput (Mbps)	End-to-End Delay (s)	Jitter (ms)	Packet Loss (%)
Background	1.68	5	440	14
Best Effort	3.63	4.26	73	25
Video	2.38	0.0012	1.1	0
Audio	0.05	0.0008	0.6	0

2) Scenario 2: The objective of this scenario 2 was to evaluate the QoS performance in handover during the implementation of ipv6 mobility support in wireless network. Based to the graph obtained in Fig. 6, the throughput of video reaches a maximum bitrate of 16 Mbps, then it settles stable at 5 Mbps during the last simulation. On the other hand, the throughput of audio takes a maximum value of 6 Mbps. The priority is given to the packets of these two access categories; this explains the straightness of the curve. The line charts of the other lower access categories (background and best effort) show an improvement and a stability in the values of productivity. As long as a high priority packet continues to send, the throughput for the background and best effort categories is lower. It increases just when the traffic for the high priority categories stops.

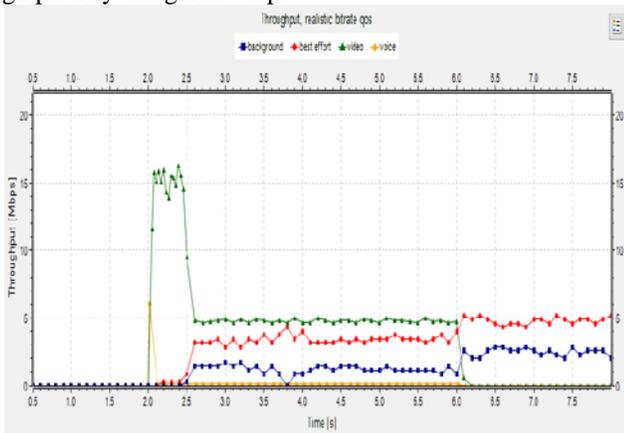


Fig. 6. Throughput variations in scenario 2

The scatter plot in Fig. 7 represents the jitter results for Scenario 2. The jitter starts at two seconds from the beginning of node mobility. The scatter graph for video and audio show horizontal data points, with values almost at zero. In contrast, the best effort and background categories showed more dispersion because the priority of these access categories is low and the packets are not sent consecutively.

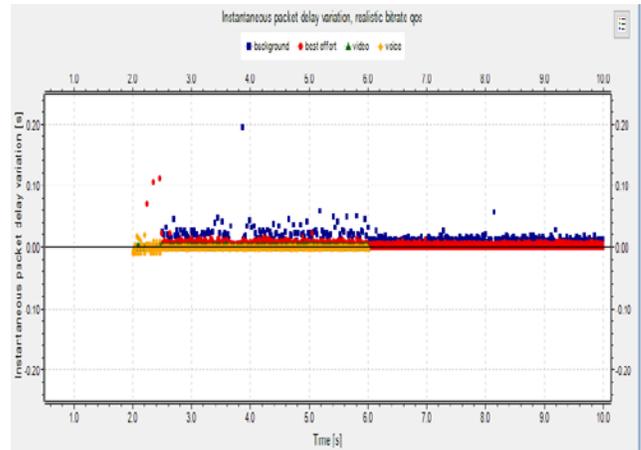


Fig. 7. Jitter variations in scenario 2

In the end-to-end delay graph shown in Fig. 8, the higher priority access categories such as video and voice are characterized by low or almost zero expected delay because they are sent before the lower priority categories like background and best effort.

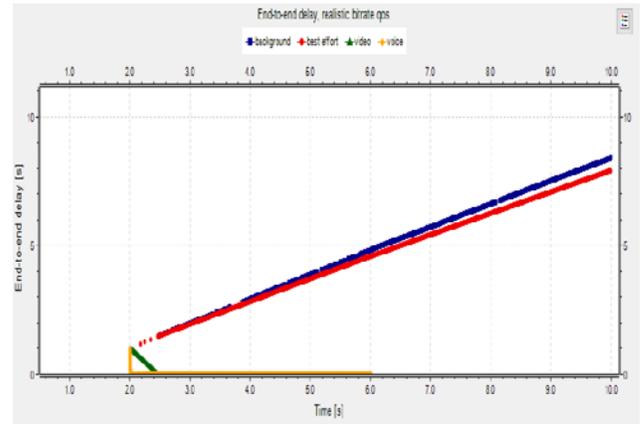


Fig. 8. End-to-end delay variations in scenario 2

Table V provides a summary of the QoS parameter values obtained for the various access categories in Scenario 2, including throughput, end-to-end delay, jitter, and packet loss:

TABLE V. SIMULATION RESULTS OF QoS IN IPV6 MOBILITY SUPPORT

Access category	Throughput (Mbps)	End-to-end delay (s)	Jitter (ms)	Packet Loss (%)
Background	1.43	5.57	380	11
Best Effort	3.12	5	67	21
Video	3.5	0.14	1.1	0
Audio	0.995	0.1	0.5	0

#### D. Simulation Comparison

Based on the results obtained, IPv6 mobility support shows relatively better performance than IPv4 mobility in terms of QoS. In particular, it produces higher average throughput values of 3.5 Mbps in video and 0.995 Mbps in audio applications, which are better than the throughput values of Mobile IPv4 as is evident from Fig. 9. The packet delay values were very low in video and audio, which distinguishes the highest priority queue with low packet delay, the background with a high value of delay because the queue priority is low, and the best effort with high throughput and medium packet delay (see Fig. 10). Regarding jitter, as seen in Fig. 11, the study shows that it is close to zero in video and audio in both Mobility IPv4 and Mobility IPv6, but more dispersed in background and best effort. Packet loss was high in background and best effort for both protocols, which explains the high level of jitter in those categories (refer Fig. 12). However, in Mobility IPv6, the packet loss values show better results compared to Mobility IPv4.

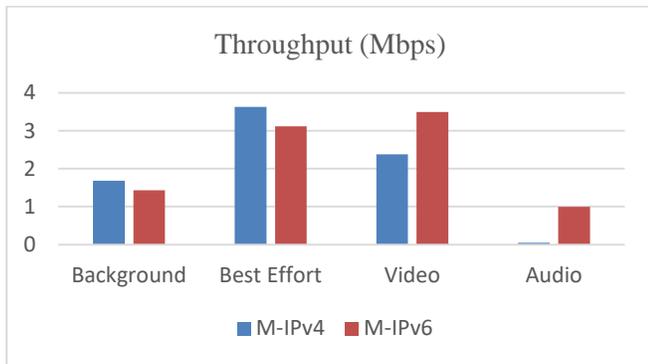


Fig. 9. Throughput comparison

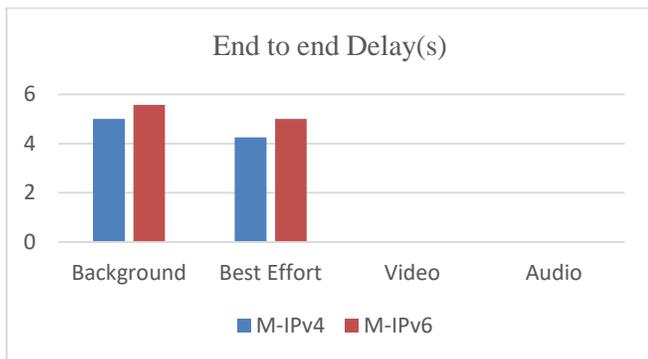


Fig. 10. End to end delay comparison

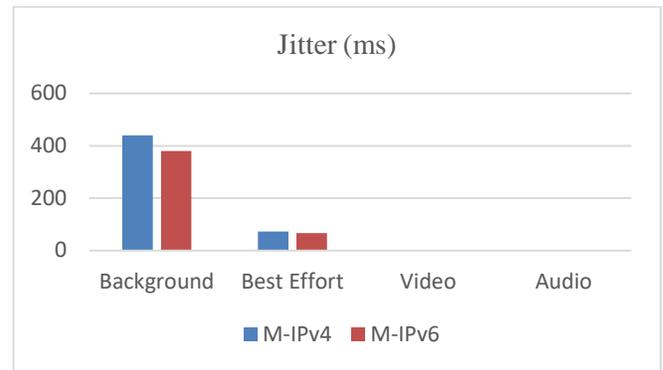


Fig. 11. Jitter comparison

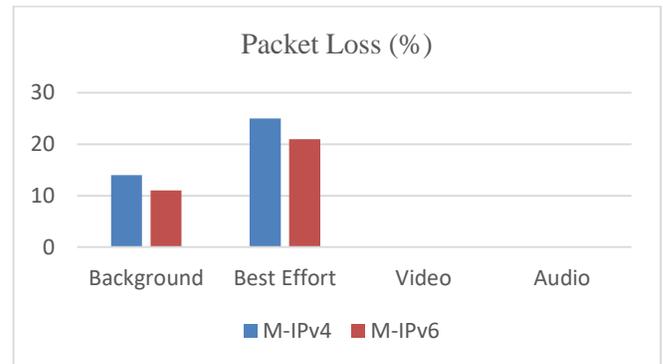


Fig. 12. Packet loss comparison

#### VI. CONCLUSION

This paper provides a detailed comparison of the QoS performance between different access categories, which is crucial for evaluating the effectiveness of the network protocols and identifying areas for improvement. By analyzing the values presented in the Table IV and V, one can gain a deeper understanding of the strengths and weaknesses of each access category in terms of QoS, allowing for more informed decision making and optimization of network performance. The study concludes IPv6 mobility support facilitates node movement in a wireless network and contributes to the improvement of quality-of-service performance. With QoS, IPv6 has a built-in mechanism for ensuring the quality of services, which makes it possible to prioritize urgent packets and to manage the processing of data packets more efficiently. In the simulation, priority is given to video and audio applications. Based on the results obtained, MIPv6 provides better QoS, with an improvement in throughput, fewer lost packets, and slight delay compared to Mobile IPv4 throughput, which was not stable in the four types of services.

## VII. FUTURE WORK

The future work could explore the potential benefits of using QoS mechanisms in MIPv4 and MIPv6 such as Differentiated Services (DiffServ) and Resource Reservation Protocol (RSVP). The results of the study could provide useful insights for the design of mobile networks that aim to provide high QoS levels for multimedia traffic.

### REFERENCES

- [1] S.Banerji, and R.S.Chowdhry, "On IEEE802.11 Wireless LAN Technology", International Journal of Mobile Network Communications & Telematics (IJMNCT), vol. 3, Issue. 4, 2013.
- [2] N.Elhilali, M.Badri, and M.Filali Bouami, "An overview of verticals handover algorithms", in press.
- [3] A.Ashraf, "A Review on Mobile Internet Protocol (Mobile IP)", International Journal for Scientific Research & Development(IJSRD), vol. 7, Issue.5, 2019.
- [4] S.K.HUSSEIN, "PERFORMANCE EVALUATION OF MOBILE INTERNET PROTOCOL VERSION 6", International Journal of Management, Information Technology and Engineering (IJMITE),pp. 35-52, vol. 4, Issue. 3, 2016.
- [5] M.Aqsa, Q.Junaid, A.Basharat, Y.Kok-Lim, and U.Ullah "QoS in IEEE 802.11-based wireless networks: A contemporary review", Journal of Network and Computer Applications, vol. 55, pp. 24-46, September 2015.
- [6] "OMNeT++ Network Simulator ", <https://omnet-tutorial.com/omnet-network-simulator/>.
- [7] E.S.Ikeremo, and Mc-Kelly.T.Pepple, "QoS PERFORMANCE EVALUATION OF VIDEO STREAMING ACROSS IPv6", RSU Journal of Biology and Applied Sciences (RSUJBAS), vol. 2, num. 1, May 2022.
- [8] A. Alwer, J.Rasheed, A.M.Abu-Mahfouz, and P.Shams" Study and Evaluation of Quality of Services in Mobile Internet Protocol v6 Using IEEE802.11e", Hindawi Wireless Communications and Mobile Computing, pp.11, November 2022.
- [9] A.Zakari,M.Musa, G.Bekaroo, S.A.Abo Bala, I.A.T.Hashem, and S.Hakak, " IPv4 and IPv6 Protocols: A Comparative Performance Study", IEEE 10th Control and System Graduate Research Colloquium, pp. 2-3, September 2019.
- [10] M.A.Al-Maqri, M.A.Alrshah, and M.Othman, "Review on QoS Provisioning Approaches for Supporting Video Traffic in IEEE802.11e: Challenges and Issues.", Digital Object Identifier, vol. 6, pp. 55202 – 55219, September 2018.
- [11] W.Zhenqi, and Y.Mei. "The Research and Application of Internet Protocol Version 6(IPv6)". Intelligent computation technology and automation (ICITA, pp.28-29 March 2011. DOI: 10.1109/ICICTA.2011.79. (Accessed: 16 August 2013).
- [12] M.Badri, M.Moughit, and N.Labani, "Association of mpls\_te and diffserv to ameliorate the QOS in mpls architecture", International Journal of Scientific and Research Publications, vol. 6, September 2016.

# Using Deep Learning Algorithms to Diagnose Coronavirus Disease (COVID-19)

Nfayel Alanazi<sup>1</sup> and Yasser Kotb<sup>2</sup>

Information Systems Department-College of Computer and Information Sciences, Imam Mohammad Ibn  
Saud Islamic University (IMSIU), Riyadh, Saudi Arabia<sup>1,2</sup>  
Computer Science Division-Department of Mathematics-Faculty of Science, Ain Shams University, Cairo,  
Egypt<sup>2</sup>

**Abstract**—With the rapid development in the area of Machine Learning (ML) and Deep learning, it is important to exploit these tools to contribute to mitigating the effects of the coronavirus pandemic. Early diagnosis of the presence of this virus in the human body can be crucially helpful to healthcare professionals. In this paper, three well-known Convolutional Neural Network deep learning algorithms (VGGNet 16, GoogleNet and ResNet50) are applied to measure their ability to distinguish COVID-19 patients from other patients and to evaluate the best performance among these algorithms with a large dataset. Two stages are conducted, the first stage with 14994 x-ray images and the second one with 33178. Each model has been applied with different batch sizes 16, 32 and 64 in each stage to measure the impact of data size and batch size factors on the accuracy results. The second stage achieved accuracy better than the first one and the 64 batch size gain best results than the 16 and 32. ResNet50 achieves a high rate of 99.31, GoogleNet model achieves 95.55, while VGG16 achieves 96.5. Ultimately, the results affect the process of expediting the diagnosis and referral of these treatable conditions, thereby facilitating earlier treatment, and resulting in improved clinical outcomes.

**Keywords**—Component; COVID-19; transfer learning; deep learning; ResNet50; VGG16; GoogleNet

## I. INTRODUCTION

In November 2019, the unique COVID-19 was reportedly discovered for the first time in Wuhan, located in Hubei Province in China. A month later, the World Health Organization (WHO) announced that the virus is capable of causing a respiratory illness that manifests clinically as coughing, fever, and inflammation of the lungs. COVID-19 was first discovered in China, but it has now been found in a significant number of different places across the globe [1, 2]. This is not only owing to the quick transmission of the virus but also the high incidence of death that has been seen as a result. Consequently, WHO declared the emergence of the new COVID-19 virus to form a pandemic [3]. WHO declared a public health emergency due to the pandemic on January 30, 2020. According to the recommendations made by the Chinese National Health Commission, radiographic evidence of pneumonia is needed as part of the clinical diagnostic criteria in Hubei Province [4], where it emphasizes the significance of different CT scan pictures for determining the degree of COVID-19 lung inflammation.

In order to a recent increase in the number of COVID-19 patients, waiting times at hospitals for CT scan image

evaluation have increased significantly. This effect causes a significant danger by spreading illness to other patients. The healthcare system becomes overburdened due to the lack of radiologists, which is generally much lower than the number of patients. In return, this causes a delay in the discovery and quarantine of infected persons in addition to the low effective treatment of patients [4]. Researchers create more intelligent, highly responsive, and efficient diagnosis approaches due to the rapid spread of COVID-19 and the enormous demand for diagnosis. In the past several years, artificial intelligence in the medical field has received widespread attention as a potentially helpful tool for guiding clinical choices and the diagnosis of diseases [5, 6]. It is noteworthy that artificial intelligence is being emphasized to work effectively in the current epidemic for the prediction of outbreaks.

A Canadian company (Blue Dot) successfully reported this outbreak's location in late December. Artificial intelligence is being emphasized to work effectively in the current epidemic to predict outbreaks. Consequently, there is an urgent need to create a clever system that can detect occurrences of COVID-19 correctly and mechanically. The research community needs to develop a dataset that is comprehensive and ready for testing as soon as possible.

CNN is the most popular machine learning algorithm which is used for image processing and training. In environment of Convolutional Neural Networks had validated very successful achievements, for example, classification of image processing, and text based, sign identification, object identifier, faces recognition. A CNN can also detect COVID traces.

This study aims to automatically identify and quantitatively evaluate the pneumonia lesions that are visible in individuals with COVID-19 who have chest CT scans. We will apply three well-known CNN models ResNet50, VGG16, and GoogleNet due to their classification efficiency in recognizing and classifying the images. The models will be compared to select the best model that can achieve high accuracy. Each model will introduce into two stages each stage has a different large dataset and different batch sizes to measure the impact and the effect of data size and batch size factors on the accuracy of the models. Each model will be trained and tested in 100 epochs. The outline of the paper is divided into the following sections. Section II illustrates the literature search. Section III discusses the proposed method in which the datasets and materials are included, where deep transfer learning models are also

explained. Section IV presents the analytical results, and finally, the paper's conclusion is drawn in Section V.

## II. RELATED WORK

There is rising interest in alternate ways of identifying coronavirus infection when using medical imaging due to the rapid spread of COVID-19. Processing and analysis of X-rays, including Computed Tomography (CT), have been performed using different deep learning methods, which have assisted physicians in predicting COVID-19 infection [7]. Several approaches that rely on identifying coronaviruses when using deep learning applications were presented. Wang and Wong [8] propose the COVID-Net approach, which represents an artificially intelligent system based on convolutional neural networks that are designed to differentiate COVID-19 instances from other types of cases by assessing lung abnormalities from X-ray pictures. A variant form of the Inception model is presented where it has an accuracy of 89.5% when extracting the features of COVID-19. This extraction uses different CT images [9]. A 3D deep learning system that is based on the location-attention mechanism is constructed, and CT scans are used in its construction. The objective of this system is to locate sick regions associated with COVID-19 patients. This method is able to differentiate COVID-19 pneumonia from influenza-related illnesses. One caused by viruses has an accuracy of 86.7% [10]. In order to identify the cases of coronavirus, a powerful type of neural network called a deep neural network is trained by using different CT scans. This network is able to differentiate among the infected areas and those affected by other lung diseases [11]. Song et al. [12] built a ResNet architecture to extract complex features from the CT data, and for the classification of COVID-19, a feature pyramid network is integrated with an attention module. In order to differentiate between instances of COVID-19 and those caused by the coronavirus, a diagnostic technique that relies on CT scans has been developed [13]. Islam et al. [14] investigate whether those who have the coronavirus can be identified by the use of chest X-rays as having the infection. The Convolutional Neural Network (CNN) and the Short-Term Long Memory (LSTM) have both been included into a new design that many researchers have constructed. The applications of machine learning, including deep learning, are conducted in COVID-19's imaging-based medical research. The primary objective of the researchers in [1] is to design a model that could make a diagnosis of COVID-19 in a manner that is analogous to the way that radiologists do it, but in a shorter period. They are able to attain performance levels that are equivalent to those of expert radiologists while reducing the amount of time spent diagnosing patients by 65% when compared to the amount of time spent by in-clinic radiologists. Despite this, there are still areas of development that can be made in order to more efficiently enhance both their suggested model and the overall system so that people may have personal access through to it. Deconstruct, transfer, and compose is the name of the deep convolutional neural network that was created and validated to recognize COVID-19 patients from the chest X-ray photos of such patients. This network is referred to by its acronym, "DeTraC" [15]. They recommended a decomposition approach to check for anomalies in the dataset by studying the class

boundary conditions. This is performed to achieve a high degree of accuracy (93.1%) and sensitivity (100%). In [16], a deep learning technique is founded on the ResNet-101 CNN model. In their suggested technique, the pre-training step consists of using thousands of photographs to distinguish significant items, and the re-training phase consists of the same images for detecting abnormalities in chest X-ray images. This approach has just a 71.9% success rate in terms of accuracy. The hospital layer, the patient layer, and the cloud layer are the proposed layers for the proposed structure, which includes three levels [17]. For the goal of collecting information from the patient layer of the data model, wearable sensors and a mobile application are utilized. An identification model based on deep learning and neural networks is used to the X-ray images of the patients in order to locate COVID-19. These images are used in conjunction with the diagnosis.

The proposed model achieves an accuracy of 97.9% and a specificity of 98.85% in its predictions. In [18], previously trained deep learning models are used, such as ResNet50, VGG16, VGG19, and DensNet121. A unique architecture is developed for the diagnosis of X-ray images as either COVID-19 or normal. The VGG16 and VGG19 models have the highest accuracy levels among those considered here. The model that is proposed consists of two phases, the first of which is preprocessing, the second of which is data augmentation, and the third of which is transfer learning. At the end, it demonstrates an accuracy of 99.3%. In the proposed model [19], three different types of deep transfer models, including AlexNet, Google Net, and ResNet18, are applied to a set of 307 photographs that include four various sorts of classes, which comprise: COVID-19, normal, pneumonia bacterial, and pneumonia viral. The models are used to classify the images. The study is broken up into three different situations to cut down on the amount of time spent when executing them and the amount of memory that is used. When it comes to the most recent deep transfer model, Google Net achieves a testing accuracy of 100% and a validation accuracy of 99.9%. In [20], a deep learning-based system is developed for identifying COVID-19 from chest X-ray images by utilizing four tuning models such as ResNet18, ResNet50, Squeeze Net, and DensNet-121. The system is able to do this by using the images as training data. The solution that is recommended involves the addition of additional data in order to generate an updated version of the COVID-19 photographs. As a direct consequence of this, the total number of samples increases. Ultimately, the obtained results achieve a sensitivity of 98% and a specificity of 90%. In [21], a model that makes use of deep learning, as well as machine learning classifiers, is developed. This model is used in a total of 38 trials to accurately identify COVID-19 through the use of chest X-ray images. Ten of these tests are carried out with a wide range of machine learning strategies, and fourteen of these experiments are carried out by utilizing a pre-trained network that is outfitted with the most recent advances in transferring learning technology. The accuracy of the system is 98.50%, while its specificity is 99.18%, and its sensitivity is 93.84%. They conclude that the CNN system can recognize COVID-19 from a limited number of photographs without the need for any preprocessing and with a decreased number of layers. This

result was obtained after they got to the opinion that the CNN system was able to.

In this research, a convolutional neural network models are presented to classify" COVID-19 images from other images, and the proposed methods are determined based on the most effective architectures, which are chosen by ILSVRC competitors as the top 5 CNN architectures. The GoogLeNet architecture won in the 2014 ILSVRC where it reduces the error rate in comparison to AlexNet and ZF-Net, and also reduces the number of parameters to 4 million in comparison to 60 million as it appears in AlexNet [22, 23]. In contrast, the VGG outperforms the other model due to the existence of the multilayer model [24] that involves nineteen more layers than the ZefNet and AlexNet. The reason behind this is to show the relations of the network representational capacity in depth. The VGG uses  $3 \times 3$  filters that are smaller than the  $11 \times 11$  and  $5 \times 5$  filters in the ZefNet. Small-size filters can produce the same impact and efficiency as large-size ones. Additionally, the small-size filter can reduce the computational complication and decrease the number of parameters. The ResNet (Residual Network) is developed in [25] and the winner of the ILSVRC in 2015, where the goal was to create an ultra-deep network to avoid the vanishing gradient issue and use the shortcut connections to enhance the deep network convergence. When compared with the VGG, the computational complexity is lower in terms of the ResNet also the enlarged depth.

In addition, many factors will affect the models leading to poor or accurate results. The study [35] aimed to point up the importance of choosing the appropriate batch size to gain the best accuracy on expected time. Different batch sizes were used to measure the best accuracy such as 16, 32, 64, and 128 they notice whenever use a small batch size that will improve the performance of the models. [36] were used 6 batch size and gained 97.78% accuracy while [39] achieved 75.51% with the same model GoogleNet, although the data size of positive COVID-19 in [36] was near to [39] but in general the number of non-COVID-19 was greater than in [36] than [39]. In the same way, the [37] and [38] use 32 batch size and [37] acquired high accuracy of 99.8% while the [38] get 76.38% with the same model ResNet50.

Consequently, some previous studies produced that the small batch size is an essential factor and has a major impact on the accuracy of the model while another study [40] gained the best accuracy by using 64 batch size. Thus, in this study, we will measure the impact of batch size and dataset size on the accuracy.

### III. PROPOSED METHOD

The proposed method is introduced to utilize three transfer learning CNN models, which comprise VGG16, ResNet50 and GoogleNet, with larger datasets from that is used in the previous studies. Different batch sizes and different images sizes are used to evaluate the performance in order to improve the results and gain the best accuracy and lower loss rate. Anaconda open-source platform is used to install and manage the Python packages which contains from 1,500+ open-source packages and that simplify the deployment and management for packages which made our experience faster and easier.

From Anaconda we create the TensorFlow environment, and we able to install all required packages on that environment.

#### A. Datasets

In this study, a publicly available accessible dataset of chest X-rays is used and taken from COVID-19 patients. These individuals either have pneumonia, normal chest X-rays or are suspected of catching COVID-19. The information for the dataset comes not just from publicly available sources, but also from hospitals and medical professionals through indirect means. The Kaggle dataset repository contains all of the photographs and data that have been made public. X-ray pictures are included in the first dataset that is collected. For the sake of this application, a COVID-19 detection model is constructed by using X-ray pictures. To create a training dataset that is divided in 70% and another scans that are divided in 30%, two datasets are used where the first of which contains 14994 images and the second which contains 33178 images are either pneumonia, normal or COVID-19 X-ray scans. The study [26] contains 2843 covid-19 images; the study [27] includes images of three categories train, test and validation. Similarly, [28] contains 2313 images of covid-19 while [29] represents many categories COVID, Lung\_Opacity, Normal and Viral Pneumonia. Both datasets are divided into 70% for the training dataset. After being rescaled, the X-ray scan pictures are shown in Fig. 2 with a size of 160.

This dataset is one that can be accessed by the general public. It presently has the most data and is labeled as of the time this article was written. In the near future, there will be a rise in the number of datasets as well as the number of samples that each one contains. Labeling is a different topic to be considered. The illness is depicted on X-ray scan pictures that are included in this collection.

#### B. Image Preprocessing

Applying the algorithms to unclear data will not give accurate and correct outcomes as it will fail to recognize the patterns effectively. Thus, data preprocessing is essential before undergoing to computations methods to enhance data accuracy.

Preprocessing is applied to the best picture of the CT Scan COVID-19 to achieve greater consistency in classification results and improved feature extraction. In the current study, assigned image resize is performed on each image to draw attention to COVID-19 in the Region of Interest (ROI), eliminate irrelevant details, and lessen the amount of work to be conducted. The CNN method requires a significant amount of iterative training. To accomplish this, a large-scale image dataset is necessary. A large-scale dataset is required to eliminate the chance of overfitting. The flow chart of the proposed approach is shown in Fig. 1 which illustrate the basic processes of our proposed study.

#### C. Data Augmentation

Several distinct data augmentation procedures are applied to the training set by utilizing the image data generator function of the Keras library, which is part of the Python programming language. This is performed to minimize overfitting and to boost the variety of the dataset. The values are by scaling, by bringing them into the same range. As a result, the rescaling

factor of 1./255 is the proposed model is applied to convert each pixel value from [0,255] to the values 0 and 1. When performing a shear transformation, one axis of the image is kept in its original position while the other axis is stretched to a predetermined angle known as the shear angle. In this example, the shear angle is set to 0.2. When performing the random zoom transformation, the zoom range argument is used; a value of less than 1.0 means that the images will be magnified. On the other hand, in order to zoom out of the image, a value that is greater than 1.0 is utilized; consequently, a zoom range of 0.2 is applied, which results in the image being magnified. The image can be flipped vertically by using the Flip function. Increase in this type of data collection.

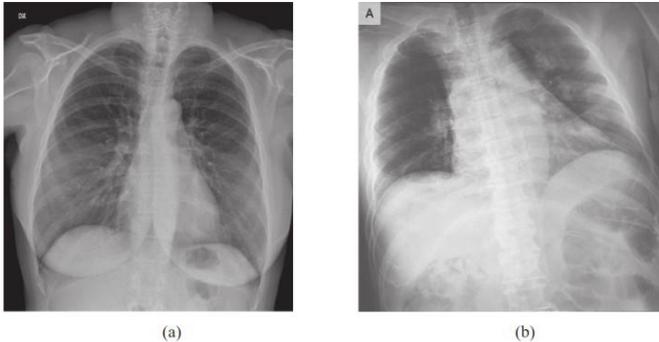


Fig. 1. Example of x-ray image taken from dataset 1

#### D. Convolutional Neural Network

In this section, we will consider the various methods of transfer learning as well as deep neural network analyses. In deep learning, the CNN neural network offers solutions in particular for the recognition, classification, and analysis of images and videos. An architecture for CNN is developed, with the visual cortex of the organization serving as a source of inspiration. Where this design is comparable to the connection model in which neurons in the brains of humans [17], CNN's recent success can be credited, at least in part, to the network's capacity to learn from large-scale datasets such as the Image Net effectively. The fundamental components of the CNN may be separated into three different levels. The neural network comprises three layers: the convolution layer, the pooling layer, and the fully connected layer. The fully connected layer is the last layer.

In conclusion, the convolutional and pooling layers are in charge of the learning that the model produces, whilst the fully connected layer is in charge of the classification [18]. The primary component of the CNN architecture is referred to as the convolutional layer. At this layer, the opportunity to acquire knowledge regarding the characteristics of the inputs is obtained. In order to produce the feature map, high-level and low-level filters are first applied to the input image. In general, the sigmoid and the ReLU function can be found in this layer.

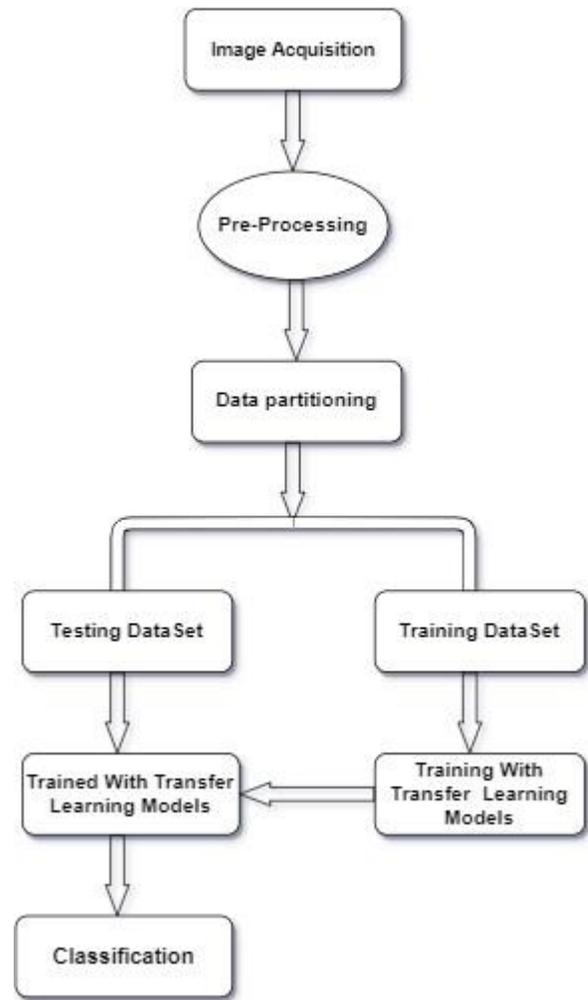


Fig. 2. The flow diagram of the proposed model

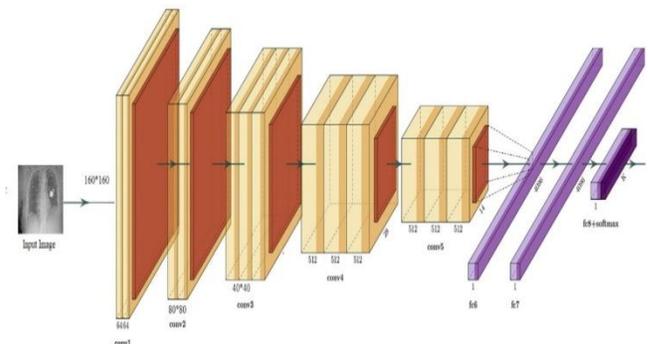


Fig. 3. The VGG 16 architecture

### E. Visual Geometry Group Network (VGG)

Simonyan et al. [30] developed the VGG-Net model, which includes minimal convolution within the network. Because of its more complex structure, which is followed by layers of related double or triple convolution layers, it is frequently used in CNN models. This is due to its structure [30]. This is the most significant difference between this model and the models that came before it, despite the fact that this model is relatively straightforward. In older models, the layers of convolution and sharing follow each other within the same order. Within this model, approximately 138 million parameters are calculated [31]. The VGG database offers an accurate representation of features for more than one million pictures (the Image Net dataset), which spans one thousand different categories. The model is able to function as an effective feature extractor for newly acquired photos that meet the requirements. The Image Net dataset has the capability to extract related features from different related photos, including entirely new ones that either does not currently exist in the dataset or that may be found in an entirely different category from those that are already existing. As a result, employing pre-trained models as an effective feature remover gives a distinct competitive edge [30]. Fig. 3 identifies the framework in which the VGG16 is built on.

Each convolution layer in the VGG16 is preceded by a ReLU layer and incorporates the maximum pooling layers for sampling. The architecture of the VGG16 employs three convolution filters and a total of 13 convolution layers to extract features. It contains three layers that are completely connected for classification, two of which a function represents hidden layers, and the final classification layer consists of one thousand units representing different picture categories that are stored in the ImageNet database [30]. Each of these layers has a different purpose. This structure provides the appearance of a larger filter while retaining the advantages of employing smaller filter sizes. It has been demonstrated that the VGGNet can function more effectively with a reduced number of parameters, particularly when compared to earlier models. Additionally, a single ReLU layer is replaced with two separate ReLU layers for the two convolution layers rather than a single ReLU layer. As a result of the convolution and partnering layers, the spatial size of the input volumes in each layer is shrunk, which leads to an increase in the volume's depth. This increment of using this depth In fact, this increment is due to the fact that the number of filters also increases. More effective when applied to object classification problems and edge detection [31].

1) *ResNet50*: ResNet50 is an architecture that is designed to have a more in-depth structure than any of the other architectures that have come before it. It is made up of 152 layers in total. In 2015, the development of the ResNet was witnessed [32]. It achieved the highest- ranking possibility in the ImageNet competition that was held in 2015, coming in first place with an error rate of 3.6% [32]. The architecture of the model's residual mapping is displayed in Fig. 4 for viewing pleasure. The blocks that feed the data to the following levels are added to the model, which is the most significant aspect that separates it from other designs. This

feature distinguishes it from other architectures. Fig. 4 illustrates the residual mapping structures where other architectures do not have this feature. The system value is altered in a manner that is described by adding this value at intervals of every two layers in the space that is occupied by the Linear and ReLU activation codes. The value of the integral from the below layer is added to the value of  $a = [Integral + 2]32$ . Increasing the number of layers in a model typically results in improved performance, but in actual practice, the situation is shifting in a different direction. In light of this, if the new theory is correct and  $w = [I + 2] = 0$ , then  $a = [I + 2] = b[I + 2]$ .

Because of this issue, the derivative will produce a value of 0 in the problem such that it is not desired [32]. However, the value feed may optimize the learning error even if the value  $a = [I]$  from the below two layers is 0. This results in the network being trained into an efficient deal more rapidly. The architecture is made up of the residual blocks that is displayed in Figure 4. The convolution of the input value  $x$  in the residual block produces  $a = F(x)$ . Fig. 4 represents the residual mapping structure.

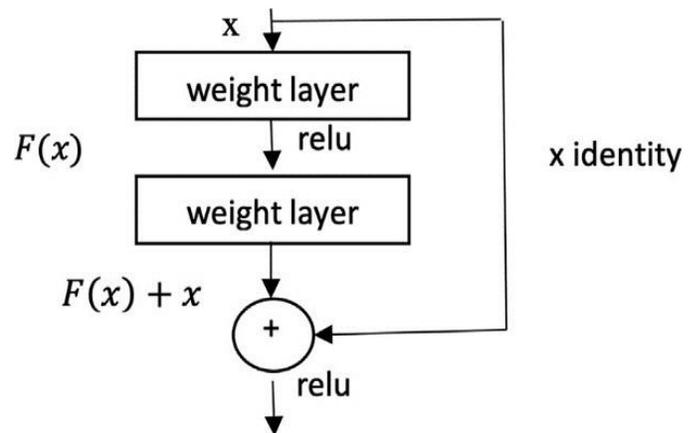


Fig. 4. Residual mapping structure

The result is applied after the ReLU convolution series. The final result is then added to the initial  $x$  entry, and the equation for  $H(x) = F(x) + x$ . Learning residuals from images rather than features is what allows the ResNet50 model to provide a simple training, including a significant advantage in its application [32].

2) *GoogleNet*: In contrast, the Google Net architecture has a total of 22 layers, making it significantly more extensive in depth and breadth than the Alex Net's architecture is, despite having a significantly lower total number of parameters in the network (five million parameters) than AlexNet has. This is due to the fact that GoogleNet has significantly fewer parameters than AlexNet has (60 million parameters). An implementation of the "network in network" design is among the most essential components of the Google Net architecture. Lin et al. [8] use a model called "inception modules." Inception makes use of parallel 1, 1, 3, 3, and 5 5 convolutions, including a parallel max-pooling layer in order to simultaneously collect a wide range of information. The

reason behind this is that it is able to record characteristics simultaneously. To satisfy the requirements of the practical application of the implementation, dimensionality reduction is accomplished by adding 1 x 1 convolutions prior to the previously mentioned 3 x 3, 5 x 5 convolutions (and also after the max-pooling layer). This is conducted in order to fulfil the expectations of the practical application. This is required because there is a requirement to maintain some level of control on the quantity of related computation. The very last layer is referred to the filter concatenation layer, and all it does is to aggregate the results of all of the layers that are running in parallel. Although this contributes to the formation of a single inception module, the version of the Google Net architecture that is utilized in the experiments that we carry out makes use of a total of nine inception modules. This is the status although the fact that this contributes to the formation of a single inception module. You can use as a reference in order to obtain a more comprehensive summary of the structure pertaining to this architecture [32].

**F. Evaluation Measures**

1) *Accuracy (ACC)*: Accuracy is the proportion of accurately predicted authentic and forged images. Accuracy is computed as:

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} * 100\% \quad (1)$$

Where True Positive (TP) refers to photos that have been accurately categorized as tampered, and False Negative (FN) refers to images that have been incorrectly labelled as tampered. True Negative (TN) denotes the photos that are initially categorized appropriately, whereas False Positive (FP) refers to the images that are first classed incorrectly. Table I presents the fundamental of evaluation metrics.

TABLE I. EVOLUTION MATRIX

		Predicted Class	
		Yes	No
Actual Class	Yes	True Positive	False Negative
	No	False Positive	True Negative

A manipulated picture is considered to have been improperly classified if it is recognized as the authentic version of the image, while an authentic image is considered to have been improperly classed if it is recognized as having been tampered with.

Here,

FN=False Negative

FP=False Positive

TN=True Negative

TP=True Positive

2) *Error*: The collective term for a model’s inaccurate predictions are known as errors. This is used to learn about all of the incorrect predictions that are produced.

$$Error = \frac{(FP + FN)}{(TP + TN + FN + FP)} * 100\% \quad (2)$$

3) *True positive rate (TPR)*: The number of fake images that are correctly discovered is known as the TPR, which is calculated as follows:

$$TPR = \frac{TP}{TP+FN} * 100\% \quad (3)$$

4) *True negative rate (TNR)*: TNR refers to the proportion of real photos that have been correctly categorized.

The following formula represents the calculation of the TNR:

$$TNR = \frac{TN}{TN+FP} * 100\% \quad (4)$$

5) *False positive rate (FPR)*: The FPR, or false positive rate, is the fraction of original photos that are incorrectly categorized. It may be computed as follows:

$$FPR = (1 - TNR) * 100\% \quad (5)$$

6) *Precision*: The total of positive predictions that are accurate is calculated as follows:

$$TPR = \frac{TP}{(TP+FN)} * 100\% \quad (6)$$

7) *Recall*: Recall is the ratio of the positive sample that is taken by model And is calculated as follows:

$$Recall = \frac{TP}{(TP+FN)} * 100\% \quad (7)$$

8) *F1-score*: The F1 score is calculated by averaging the recall of the model with the accuracy of the model. One is maximal and zero is minimal value. If the value is at its highest possible, the model is said to be of high quality.

$$F1 - score = \frac{2 * (Precision * Recall)}{(Precision + Recall)} * 100\% \quad (8)$$

9) *F-Measure*

$$F = \frac{2TP}{(2TP+FP+FN)} \quad (9)$$

IV. RESULTS AND DISCUSSION

TABLE II. EXPERIMENT RESULT

Model Name	Batch Size	Epoch Size	Testing Accuracy	Dataset Size
GoogleNet	16	30	94	14994
	32	21	94.46	
	64	34	93.64	
ResNet50	16	91	94.84	
	32	88	96.11	
	64	84	93.3	
VGG16	16	61	93.55	
	32	89	94	
	64	41	93.3	
GoogleNet	64	84	95.62	33178
ResNet50	128	86	96.07	
	64	77	96.07	
VGG16	64	89	95.81	

A. Feature Extraction Performance

The trained CNN is in its system on extracted characteristics before using it to categorize the data. To assess the effectiveness of the CNN models, test features are extracted from the test pictures and are analyzed with a variety of pre-trained models. Currently, the CNN models extract features from both the training and testing datasets using various pre-trained models, such as Google Net, ResNet50, VGG-16, and ResNet50. In the current study, a comparison of the performance of several CNN models is carried out. Even though the ResNet50 does not demonstrate the highest performance in terms of greater accuracy, it is clear that the ResNet50, which is suggested in this study, obtains a superior accuracy and specificity than the other pre-trained models. However, when compared to CNN models, the performance of the scratch model is not considered to be sufficient. The results of the produced ResNet50 model are much superior to those obtained by the VGG16 and GoogleNet models presented in Table II.

B. Comparative Analysis

It is a time-consuming process to use the appropriate datasets of chest X-ray pictures for the COVID-19 detection process. The researchers made use of a variety of preprocessing methods, feature extraction strategies, and classification approaches [24, 33].

However, it is difficult to identify a prospective strategy or combination of techniques that are more supportive in diagnosing COVID-19 from the chest X-ray picture. The reason behind this is that there are so many different possible approaches. In the vast majority of instances, a level of accuracy that is found greater than 90% is observed from a statistical perspective, this constitutes a very high degree of accuracy. However, the objective would be to improve the degree of accuracy to be as near as possible to 100%, where it is given that incorrect diagnosis, even in a very small number of instances, is not quite acceptable. It is very obvious that the approach that is proposed creates a greater classification accuracy when it comes to identifying COVID-19 in comparison to the other strategies that are proposed in the literature. However, the study by Loye et al. [19] shows a more accurate result of 100% than what is found in this study. This might be because the dataset that they use to assess the performance of the system only have a very small number of photos (69 COVID-19 and 79 normal images, respectively) [34]. The accuracy of the suggested technique, which utilizes the feature fusion that is derived from Googlenet, ResNet50, and CNN (VGG16), is demonstrated to be higher when utilizing the CNN as the classifier. On the other hand, when using a binary classifier on the chest X-ray dataset, the ResNet50 achieves a satisfactory level of performance.

In Table II, a different experiment is illustrated when using different batch sizes and epochs when also using the 14994 images dataset. The first set of experiments with GoogleNet employs three different batch sizes (16, 32, and 64) and epochs (30, 21, and 34), and the results show that the testing accuracy reaches 94%, 94.46%, and 93.64% as shown in graph in Fig. 5.

The second ResNet50 model is utilized for experimental purposes, and it achieves a testing accuracy of 94.84%, 96.11%, and 94.53% by utilizing the same batch sizes 16, 32, 64 and epochs used in 91, 88, and 83, correspondingly as shown in graph Fig. 6. The final experiment in the 14994-image dataset is performed by using the VGG16 model, which achieves 93.55%, 93.75%, and 94% as shown in graph Fig. 7, respectively, when using the same batch size 16, 32, 64 and epochs used in 61,89,41. Following that, an experiment is performed with X-ray images by using an increased dataset of 33178 images. When using the same model three times, the batch size and time period each time are changed. The very first GoogleNet model includes a testing accuracy of 95.64% and uses batch sizes of 64 and 84, respectively. The second ResNet50 model should then be run with two different batch sizes of 64 and 128, while maintaining the same level of testing accuracy, which should be 96.07%. The final VGG16 model that is utilized for the experiment, which includes a batch size of 64, and its epoch's size reaches 89. It is able to reach an accuracy of 95.81%. Both the VGG16 and GoogleNet models almost achieve the same level of performance when it comes to the COVID-19 classification. The ResNet50 model, which provides the maximum classification performance with a testing accuracy of 95%, is used for both datasets in order to conduct the evaluation. 32, 64, and 128 batch sizes are applied in order to measure the accuracy in each batch and select the best batch size.

Regardless of the fact that the picture size is decreased by more than half of the original sizes in this experiment, the system is still exhibiting its resilience by accurately detecting COVID-19 instances. This is the case although the fact that the image size is lowered. One possible explanation for this is because the state-of-the-art algorithms have been able to identify a greater variety of unique characteristics and extract those characteristics.

Fig. 8 demonstrates confusion matrix of ResNet50 in first stage, Fig. 9 represents confusion matrix of VGG16 also in first stage while Fig. 10 shows the confusion matrix of VGG16 and ResNet50 in second stage.

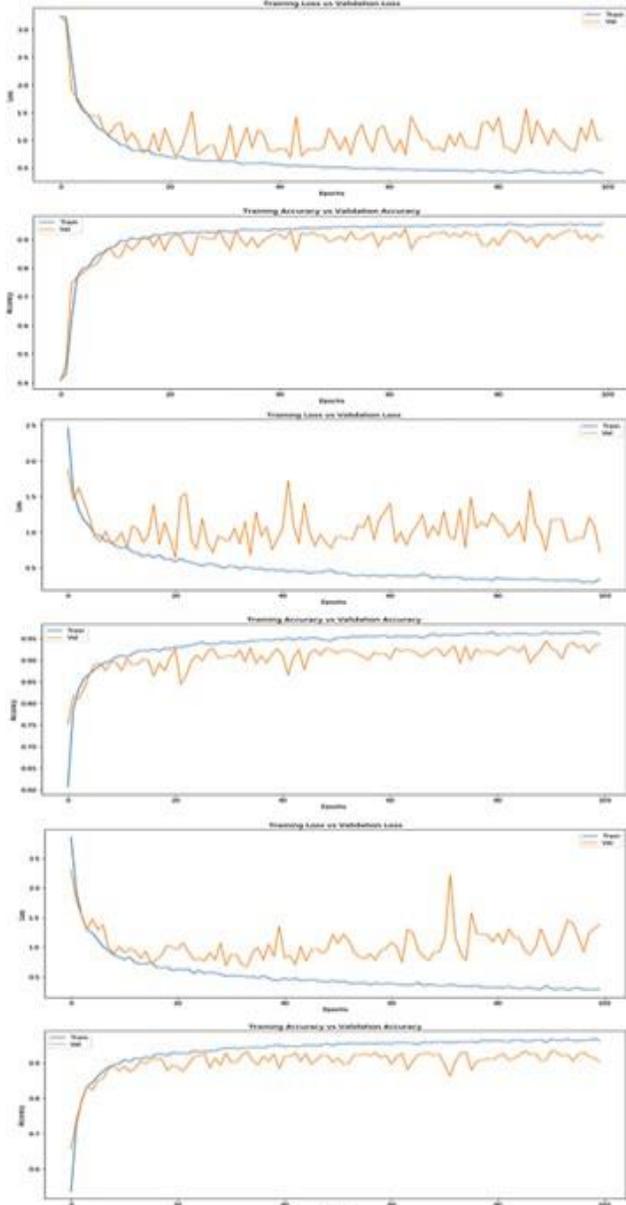


Fig. 5. Accuracies and loss error rates results graphs of proposed GoogleNet model in Dataset 1

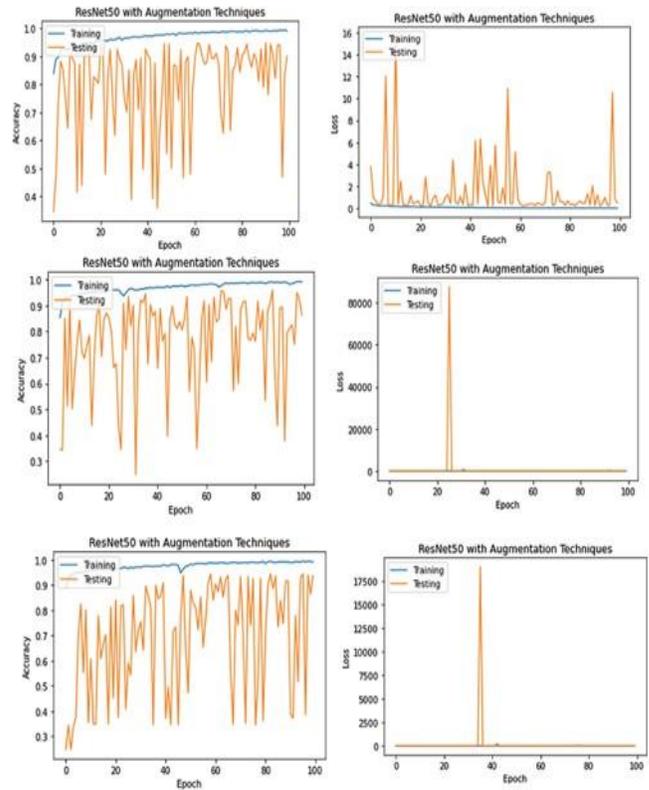


Fig. 6. Accuracies and loss error rates results graphs of proposed ResNet50 model in Dataset 1

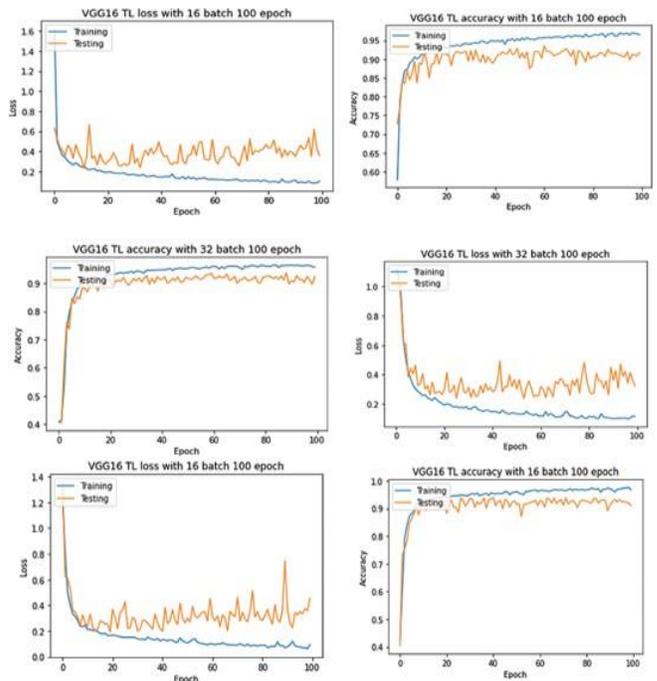


Fig. 7. Accuracies and loss error rates results graphs of proposed VGG16 model in Dataset 1

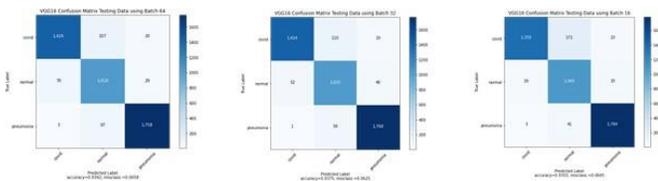


Fig. 8. ResNet50 model confusion matrix

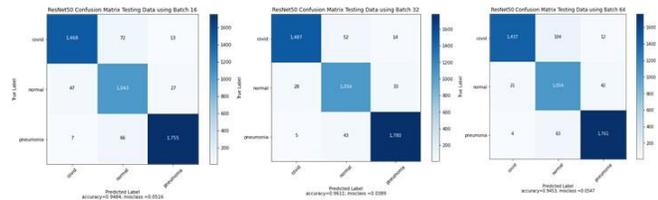


Fig. 9. VGG16 model confusion matrix

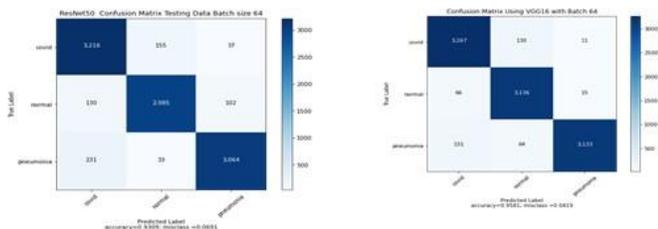


Fig. 10. ResNet50 and Vgg16 confusion matrix big dataset

## V. CONCLUSION

The rationale behind utilizing X-ray pictures in the COVID-19 detection process is initially discussed in this paper. After that, a few similar papers on pre-trained CNN systems utilizing X-ray pictures are highlighted. In order to create a COVID-19 detection, a large public dataset consisting of chest X-rays is used because there is insufficient public COVID-19 datasets. The first dataset contains 14994 COVID-19 X-ray pictures, while the second dataset contains 33174 COVID-19 X-ray. These are used for training and testing stages, respectively. The photographs are scaled to be 160 by 160 pixels. The dataset is made more comprehensive by including different picture enhancement techniques.

Transfer learning models GoogleNet, ResNet50, and VGG16 were used by utilizing the X-ray images of COVID-19 patients and other patients. The pre-trained ResNet50 model produced the highest classification performance of automated COVID-19 classification with an accuracy of 96%. This was compared to the other two suggested models, which both had a classification accuracy of 94%. By presenting our work's results in graphs and tables, we were able to draw attention to the performance of the categorization. It is recommended that the number of examples that are included in the dataset to be raised so that the model can achieve a more accurate performance. In addition, we hope in the future to use pre-processing techniques that can affect positively on the results and that will increase the intelligence of the models and will lead to predicting the accurate result that can be generalized.

## REFERENCES

[1] Chen, J.; Wu, L.; Zhang, J.; Zhang, L.; Gong, D.; Zhao, Y.; Chen, Q.; Huang, S.; Yang, M.; Yang, X.; others. Deep learning-based

model for detecting 2019 novel coronavirus pneumonia on high-resolution computed tomography. *Scientific reports* **2020**, *10*, 1–11.

[2] Holshue, M.L.; DeBolt, C.; Lindquist, S.; Lofy, K.H.; Wiesman, J.; Bruce, H.; Spitters, C.; Ericson, K.; Wilkerson, S.; Tural, A.; others. First case of 2019 novel coronavirus in the United States. *New England journal of medicine* **2020**.

[3] Ismail, W.N.; Hassan, M.M.; Alsalamah, H.A.; Fortino, G. CNN-based health model for regular health factors analysis in internet-of-medical things environment. *IEEE Access* **2020**, *8*, 52541–52549.

[4] Maghded, H.S.; Ghafoor, K.Z.; Sadiq, A.S.; Curran, K.; Rawat, D.B.; Rabie, K. A novel AI-enabled framework to diagnose coronavirus COVID-19 using smartphone embedded sensors: design study. 2020 IEEE 21st International Conference on Information Reuse and Integration for Data Science (IRI). IEEE, 2020, pp. 180–187.

[5] Topol, E.J. High-performance medicine: the convergence of human and artificial intelligence. *Nature medicine* **2019**, *25*, 44–56.

[6] Tomašev, N.; Glorot, X.; Rae, J.W.; Zielinski, M.; Askham, H.; Saraiva, A.; Mottram, A.; Meyer, C.; Ravuri, S.; Protsyuk, I.; others. A clinically applicable approach to continuous prediction of future acute kidney injury. *Nature* **2019**, *572*, 116–119

[7] Wu, X.; Hui, H.; Niu, M.; Li, L.; Wang, L.; He, B.; Yang, X.; Li, L.; Li, H.; Tian, J.; others. Deep learning-based multi-view fusion model for screening 2019 novel coronavirus pneumonia: a multicentre study. *European Journal of Radiology* **2020**, *128*, 109041.

[8] Wang, L.; Lin, Z.Q.; Wong, A. Covid-net: A tailored deep convolutional neural network design for detection of covid-19 cases from chest x-ray images. *Scientific Reports* **2020**, *10*, 1–12.

[9] Wang, S.; Kang, B.; Ma, J.; Zeng, X.; Xiao, M.; Guo, J.; Cai, M.; Yang, J.; Li, Y.; Meng, X.; others. A deep learning algorithm using CT images to screen for Corona Virus Disease (COVID-19). *European radiology* **2021**, *31*, 6096–6104.

[10] Hou, H.; Lv, W.; Tao, Q. Hospital, T., Company, JT, Ai, T. *Hospital, T., Wuhan, T., & Hospital* **2020**.

[11] Cheng Jin, J.; Chen, W.; Cao, Y.; Zhanwei, X.; Zhang, X.; Deng, L.; Zheng, C.; Zhou, J.; Shi, H. Development and evaluation of an AI system for COVID-19 diagnosis. *MedRxiv* **2020**.

[12] Song, Y.; Zheng, S.; Li, L.; Zhang, X.; Zhang, X.; Huang, Z.; Chen, J.; Wang, R.; Zhao, H.; Chong, Y.; others. Deep learning enables accurate diagnosis of novel coronavirus (COVID-19) with CT images. *IEEE/ACM transactions on computational biology and bioinformatics* **2021**, *18*, 2775–2780.

[13] Takahashi, M.S.; de Mendonça, M.R.F.; Pan, I.; Pinetti, R.Z.; Kitamura, F.C. Regarding serial quantitative chest CT assessment of COVID-19: deep-learning approach. *Radiology: Cardiothoracic Imaging* **2020**, *2*.

[14] Islam, M.Z.; Islam, M.M.; Asraf, A. A combined deep CNN-LSTM network for the detection of novel coronavirus (COVID-19) using X-ray images. *Informatics in medicine unlocked* **2020**, *20*, 100412.

[15] Abbas, A.; Abdelsamea, M.M.; Gaber, M.M. Classification of COVID-19 in chest X-ray images using DeTraC deep convolutional neural network. *Applied Intelligence* **2021**, *51*, 854–864.

[16] Che Azemin, M.Z.; Hassan, R.; Mohd Tamrin, M.I.; Md Ali, M.A. COVID-19 deep learning prediction model using publicly available radiologist-adjudicated chest X-ray images as training data: preliminary findings. *International Journal of Biomedical Imaging* **2020**, *2020*.

[17] El-Rashidy, N.; El-Sappagh, S.; Islam, S.R.; El-Bakry, H.M.; Abdelrazek, S. End-to-end deep learning framework for coronavirus (COVID-19) detection and monitoring. *Electronics* **2020**, *9*, 1439.

[18] Khan, I.U.; Aslam, N. A deep-learning-based framework for automated diagnosis of COVID-19 using X-ray images. *Information* **2020**, *11*, 419.

[19] Loey, M.; Smarandache, F.; M. Khalifa, N.E. Within the lack of chest COVID-19 X-ray dataset: a novel detection model based on GAN and deep transfer learning. *Symmetry* **2020**, *12*, 651.

- [20] Minaee, S.; Kafieh, R.; Sonka, M.; Yazdani, S.; Soufi, G.J. Deep-COVID: Predicting COVID-19 from chest X-ray images using deep transfer learning. *Medical image analysis* **2020**, *65*, 101794.
- [21] Sekeroglu, B.; Ozsahin, I. <? covid19?> Detection of COVID-19 from Chest X-Ray Images Using Convolutional Neural Networks. *SLAS TECHNOLOGY: Translating Life Sciences Innovation* **2020**, *25*, 553–565.
- [22] Iandola, F.N.; Han, S.; Moskewicz, M.W.; Ashraf, K.; Dally, W.J.; Keutzer, K. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and < 0.5 MB model size. *arXiv preprint arXiv:1602.07360* **2016**.
- [23] Alom, M.Z.; Taha, T.M.; Yakopcic, C.; Westberg, S.; Sidike, P.; Nasrin, M.S.; Van Esesn, B.C.; Awwal, A.A.S.; Asari, V.K. The history began from alexnet: A comprehensive survey on deep learning approaches. *arXiv preprint arXiv:1803.01164* **2018**.
- [24] Ahammed, K.; Satu, M.S.; Abedin, M.Z.; Rahaman, M.A.; Islam, S. Early detection of coronavirus cases using chest X-ray images employing machine learning and deep learning approaches. *MedRxiv* **2020**.
- [25] Vedaldi, A.; Lenc, K. Matconvnet: Convolutional neural networks for matlab. *Proceedings of the 23rd ACM international conference on Multimedia*, 2015, pp. 689–692.
- [26] Hussain, E. (2020) Covid R - largest Covid 19 dataset, Kaggle. Available at: <https://www.kaggle.com/ehrupok/covid-r-largest-covid-19-dataset?select=Covid>.
- [27] Mooney, P. (2018) Chest X-ray images (pneumonia), Kaggle. Available at: <https://www.kaggle.com/datasets/paultimothymooney/chest-xray-pneumonia>.
- [28] Asraf, A. (2020) Covid19\_pneumonia\_normal\_chest\_xray\_pa\_dataset, Kaggle. Available at: <https://www.kaggle.com/amanullahasraf/covid19-pneumonia-normal-chest-xray-pa-dataset?select=covid>.
- [29] Rahman, T. (2022) Covid-19 radiography database, Kaggle. Available at: <https://www.kaggle.com/datasets/tawsifurrahman/covid19-radiography-database>.
- [30] Guo, T.; Dong, J.; Li, H.; Gao, Y. Simple convolutional neural network on image classification. 2017 IEEE 2nd International Conference on Big Data Analysis (ICBDA). IEEE, 2017, pp. 721–724.
- [31] Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* 2014.
- [32] Zeiler, M.D.; Fergus, R. Visualizing and understanding convolutional networks. *European conference on computer vision*. Springer, 2014, pp. 818–833.
- [33] Saparkhojayevev, N.; Zholayeva, L.; Tashkenbayev, Y.; Tokseit, D. Abnormality Detection in Chest X-ray Images Using Uncertainty Prediction Algorithms. 2021 16th International Conference on Electronics Computer and Computation (ICECCO). IEEE, 2021, pp. 1–3.
- [34] Yasar, H.; Ceylan, M. A new deep learning pipeline to detect Covid-19 on chest X-ray images using local binary pattern, dual tree complex wavelet transform and convolutional neural networks. *Applied Intelligence* **2021**, *51*, 2740–27.
- [35] Darapaneni, N., Ranjane, S., Satya, U. S. P., Reddy, M. H., Paduri, A. R., Adhi, A. K., et al. (2020). "COVID-19 severity of pneumonia analysis using chest x rays," in 2020 IEEE 15th International Conference on Industrial and Information Systems (ICIIS) (IIT Ropar), 381–386. doi: 10.1109/ICIIS51140.2020.9342702.
- [36] Misra, S., et al. (2020). "Multi-channel transfer learning of chest X-ray images for screening of COVID-19." *Electronics* 9(9): 1388.
- [37] Miroshnichenko, A. and V. Mikhelev (2021). Classification of medical images of patients with Covid-19 using transfer learning technology of convolutional neural network. *Journal of Physics: Conference Series*, IOP Publishing.
- [38] Mohammadi, R., et al. (2020). "Transfer learning-based automatic detection of coronavirus disease 2019 (COVID-19) from chest X-ray images." *Journal of Biomedical Physics & Engineering* 10(5): 559.
- [39] El Gannour, O., Hamida, S., Cherradi, B., Raihani, A., and Moujahid, H. (2020). "Performance evaluation of transfer learning technique for automatic detection of patients with COVID-19 on X-Ray images," in 2020 IEEE 2nd International Conference on Electronics, Control, Optimization and Computer Science (ICECOCS) (Kenitra), 1–6. doi: 10.1109/ICECOCS50124.2020.9314458.
- [40] Bhatia, N. and G. Bholra (2021). Transfer learning for detection of COVID-19 infection using chest X-ray images. 2021 5th International Conference on Computing Methodologies and Communication (ICCMC), IEEE.

# Enhanced Optimized Classification Model of Chronic Kidney Disease

Shahinda Elkholy<sup>1</sup>, Amira Rezk<sup>2</sup>, Ahmed Abo El Fetoh Saleh<sup>3</sup>

Information Systems Dept.-Faculty of Computers and Information, Mansoura University,  
Mansoura, Egypt

**Abstract**—Chronic kidney disease (CKD) is one of the leading causes of death across the globe, affecting about 10% of the world's adult population. Kidney disease affects the proper function of the kidneys. As the number of people with chronic kidney disease (CKD) rises, it is becoming increasingly important to have accurate methods for detecting CKD at an early stage. Developing a mechanism for detecting chronic kidney disease is the study's main contribution to knowledge. In this study, preventive interventions for CKD can be explored using machine learning techniques (ML). The Optimized deep belief network (DBN) based on Grasshopper's Optimization Algorithm (GOA) classifier with prior Density-based Feature Selection (DFS) algorithm for chronic kidney disease is described in this study, which is called "DFS-ODBN." Prior to the DBN classifier, whose parameters are optimized using GOA, the proposed method eliminates redundant or irrelevant dimensions using DFS. The proposed DFS-ODBN framework consists of three phases, preprocessing, feature selection, and classification phases. Using CKD datasets, the suggested approach is also tested, and the performance is evaluated using several assessment metrics. Optimized-DBN achieves its maximum performance in terms of sensitivity, accuracy, and specificity, the proposed DFS-ODBN demonstrated accuracy of 99.75 percent using fewer features comparing with other techniques.

**Keywords**—Machine learning (ML); feature selection (FS); chronic kidney disease (CKD); deep belief network (DBN); grasshopper's optimization algorithm (GOA)

## I. INTRODUCTION

Chronic renal disease, or chronic kidney disease (CKD), worsens over time, eventually causing kidney failure. Most of the time, it can go undetected for up to 25% of its usefulness before being discovered. For people who are unaware of kidney failure's symptoms, this might make it difficult to diagnose and treat the condition. Treatment for kidney failure aims to manage the causes and slow the progression of renal failure. Patients in the last stages of renal failure will require dialysis or a kidney transplant if treatment fails [1]. Renal failure affects four out of every 1000 people in the UK, while dialysis keeps more than 300,000 Americans with kidney failure alive [2]. More people in South Asia, Africa, and the rest of the world have renal illness, according to the National Health Service (NHS). Because chronic renal failure cannot be detected until it has progressed to an advanced state, recognizing kidney failure at an early stage is critical. By detecting renal disease at an early stage, the likelihood of permanent damage to the kidneys is reduced. As a result, patients should have regular checkups and early diagnosis to avoid serious risks of renal failure and related disorders [1].

Doctors can make therapy decisions that slow the rate of progression. Measuring parameters allow for differentiation, and patients' medical records can be used to classify and predict disease using data mining techniques [3].

Data mining techniques allow for the extraction of meaningful information from large and hidden databases. As a method of gaining knowledge from unstructured information, data mining techniques can be applied even when the information is not directly related to medicine [4]. Data mining has three stages: data processing, data modeling, and data post processing. Data mining jobs in data modeling include classification/predictive algorithms and regression algorithms that are learned through a supervised learning process. As a result of missing and unneeded data being stored in the hospital database, it is difficult to mine the patient data. As a result, prior to implementing data mining techniques, it is necessary to improve data processing and data reduction methodologies [5]. Accurate and reliable data makes the identification of CKD easier and faster. Data classification can be used to identify CKD from a patient's medical records but an important part of the classification process is the establishment of a link between the feature values and the class labels for the data being processed. Hence, classification is a supervised procedure. An algorithm for classification and prediction uses training data to build a model, which is then used to predict test data [6]. Using artificial intelligence (AI) approaches, categorization models have recently been improved. Multiple issues arose as a result of the high-dimensional nature of the medical data, including high processing complexity, overfitting, and low finishing model interoperability. Feature selection (FS) is the quickest and most effective way to address the issue. This method aims to reduce the number of features to a manageable subset by removing redundant or irrelevant ones. In order to save calculation time, it makes use of a small number of characteristics to extract the maximum amount of data from a dataset [7]. The selected feature subset is useful in modeling these functions. To improve prediction results, FS approaches are used in a wide range of applications such as machine learning (ML), data mining, and pattern recognition [8]. Wrapper, embedded, and filter-based methods are all examples of FS validation methods. To validate the feature subset in a filter technique, one uses fixed measurements rather than learners and a predetermined set of features. However, the wrapper method uses the learning strategy as a sub-process of evaluation to determine whether or not a feature set is better than it was previously. This method, which is widely used, has certain drawbacks, such as a high computational cost, difficulty in recognizing user-defined parameters of the learner, and

built-in constraints on the learners [9]. Embedded techniques are simpler than wrapper approaches in terms of complexity. But the selection of features is based on the learning technique [10]. The filter and wrapper methods are both integrated into the embedding method, which removes their respective drawbacks. Despite the cheap computational complexity of the filter approaches, the feature subset used for classification was shown to be unreliable. The wrapper approaches, on the other hand, achieve better classification results while using a lot more time. All three techniques have improved the classification of the features. In addition, the FS procedure has improved features rather than the classifier. High classification performance was achieved with high computing complexity using wrapper and hybrid approaches.

This paper aims to enhance the accuracy of CKD classification using the DFS-ODBN method which an innovative wrapper strategy for CKD detection that incorporates density-based feature selection (DFS) with optimized DBN based on GOA to tackle these difficulties. Used techniques are capable of solving the problem of class imbalance in the dataset which may affect classification performance in addition to this heuristic way that determine whether or not a feature is worthwhile. To improve the final classifier outputs of the DFS-ODBN algorithm, the addition of optimization algorithm called GOA as parameter adaptation to increase the performance of the classifier helps to enhance the accuracy of the optimized DBN classifier. From the UCI repository, a benchmark CKD dataset is used to test the efficiency of the DFS-ODBN method. The model was evaluated using metrics like accuracy, sensitivity, and F-measure. The results showed that the provided DFS-ODBN strategy outperforms the evaluated approaches in terms of classification performance.

The following is the article's flow: Other diagnostic approaches are reviewed in Section II; Section III presents a review of related methods and technologies; Section IV provides a discussion of applied methodology, and Section V includes simulation test results for the proposed method, comparing the results with other related studies. Section VI provides the future work with the concluded presented work.

## II. RELATED WORK

Much research has used data mining algorithms to accurately predict CKD in patients based on their medical records. From the original set of features, there has been a focus on a subset of relevant features that play a significant role in the medical diagnosis sectors because of the high dimension of required multidimensional medical multimedia data for CKD prediction. All of the studies used performance indicators including specificity, accuracy, and sensitivity to support their CKD prediction approach. We'll go through some current research on predicting kidney disease later on.

The UBFST (Union Based Feature Selection Technique) was developed for the rapid and accurate classification and diagnosis of chronic kidney disease (CKD). This method uses SVM, regression tree, and random forest to classify CKD [11]. Using a Las Vegas Wrapper Feature Selection approach (LVW-FS) and an ensemble learning-based model, hemodialysis treatment time can be predicted with acceptable

accuracy. Using the LVW selection methodology, they have developed a new way of extracting crucial vital indicators. As a method of classification, a group of learners was used in this study to give numerous classifiers. Through a variety of trials with different learners, the suggested model based on LVW and the ensemble learning method was shown to have the greatest influence in decreasing hyperthyroidism characteristics and excluding noise [12]. Their technology can forecast the health of the kidney based on factors such as age, albumin and glucose levels, and more.

An evolutionary algorithm (GA) based on neural networks optimize weight vectors to train a neural network. For CKD diagnosis, the system outperforms existing neural networks in terms of accuracy [13]. For the dataset of CKD, using multilayer perceptrons (MLPs), probabilistic neural networks (PNNs), radial basis functions (RBFs), and SVM. The PNN algorithm surpassed the SVM, MLP, and RBF algorithms in terms of performance [14]. In Colombian population neural networks were used to predict the likelihood of CKD in people [15].

The study [16] proposed a method for diagnosing chronic renal illness based on grey wolf optimization (GWO) and hybrid kernel support vector machines (HKSVM). The UCI ML repository's chronic kidney dataset yielded a 97.26% accuracy rate. CKD can be identified using two fuzzy classifiers known as FuRES and FOAM, which are both fuzzy rule-building expert systems (FuRES). FuRES provides a minimum NN-based classification tree. The weight vector with the least fuzzy entropy is determined by the categorization criteria. The 386 CKD patients were identified using two fuzzy classifiers. FuRES, on the other hand, performs better than FOAM in cases where the training and prediction processes are both noisy. In the detection of CKD, both FOAM and FuRES performed better, although FuRES outperformed FOAM.

PCA and SVM were used to diagnose cervical cancer. A total of 32 potential risk factors as well as four specific outcomes were examined in this study: Hinselmann, Schiller, cytology, and biopsies. SVM with recursive feature removal and SVM with PCA were used to classify the target objects (PCA-SVM). PCA-SVM came out on top over the other two methods [17].

Using three classification algorithms Olex-GA, Ant Colony Optimization ACO and PSO compared them to developed system for diagnosing CKD. The ACO method with Density Based Feature Selection to select the important features showed superior performance and the highest accuracy [18].

Seven ML techniques are comprised to predict the CKD, J48, SVM, NBTree, LR, MLP, Naïve Bayes and Composite Hypercube on Iterated Random Projection (CHIRP) are utilized. The results of experiments show better performance for CHIRP [19].

C5.0, Artificial neural network, CHAID, logistic regression, random tree, K-Nearest neighbors and linear support vector machine were the seven classifier algorithms used to predict CKD. Results were calculated using all features from the classifier, features chosen by CFS, features chosen by Wrapper, LASSO regression, SMOTE, and chosen features

SMOTE with all the features, LASSO with selected features. The LSVM was found to have the highest accuracy in SMOTE with all features [20].

The research [21], has an additional strategy based on Recursive Feature Elimination (RFE). The most strongly representative features of CKD were chosen using the RFE algorithm. SVM, KNN, decision tree, and random forest were used to classify the features. All classifier parameters were fine-tuned to provide the best classification results, and all methods produced promising results. For all measures, the random forest approach surpassed all other algorithms. Multiclass statistical analysis was used to investigate and evaluate the system, and the empirical results of SVM, KNN, and decision tree algorithms revealed significant values of 96.67 percent, 98.33 percent, and 99.17 percent in terms of accuracy metric.

To properly pick the features subset, several feature optimization techniques were described to see the effect of them on the performance of the ML model which was tested on five influential classification models Logistic regression, Random forest (RF), SVM, K-nearest neighbors and Xtreme gradient boosting (XGB), experiments have shown that the accuracy of the model can be enhanced by using Linear discriminant analysis (LDA) feature optimizer that performs the highest outstanding result [22].

### III. RELATED METHODOLOGIES

#### A. Feature Selection

Pattern recognition, knowledge discovery, and statistical research all rely heavily on feature selection. It is the goal of feature selection to eliminate unneeded inputs. No predictive class information is required to determine which features are relevant. Reducing the dimensionality of features and omitting features that are not relevant to classification can result in a comprehensive model. As the name implies, the fundamental issue in feature reduction lies in identifying the optimal collection of features to maximize classification performance [23]. Simplifying the data collection, reducing the problem of over fitting, and reducing the amount of data stored are all benefits of feature selection [24].

Three types of feature selection approaches exist. A variety of methods, including filtering, embedding, and wrapping methods. The filter method chooses the highest-ranking features, and the resulting subset can then be used in any classification algorithm that is needed to be used. Various classification methods could be tested after feature selection using the filter approach as shown in Fig. 1 [25]. The classifier's performance can be improved by reducing processing time and making better use of the dataset's optimized data while making a suitable feature selection [26]. Fastness and scalability are two more advantages of the filter approach in feature selection [27].

The classifier algorithm is used as a black box to determine scores for feature sets based on estimated power [28]. Testing and training on a given dataset are used to evaluate a subset. All features of subsets can be searched for in a wrapper approach that uses the wrapper algorithm around the classifier shown in Fig. 2 [25]. In spite of the advantages of working with correlated data and identifying the relevant correlations, over-fitting difficulties may arise. Feature selection is included in the construction of the classifier in the embedded method illustrated in Fig. 3.

When using an embedded technique, you don't have to deal with costly computations and can instead interact directly with your classification model. Table I displays the benefits and drawbacks of various feature selection approaches.

TABLE I. BENEFITS AND DRAWBACKS OF VARIOUS FEATURE SELECTION APPROACHES

approach	Benefits	Drawbacks
<b>Filter</b>	Scalable and quick without being dependent on any classifier,	Interaction with classifiers is not considered.
<b>Wrapper</b>	- A higher degree of computational efficiency, Simple, Classification accuracy models with dependencies are the best kind. - The classifier interacts with it. - Wrapper approaches have a high computational complexity. Models include a hierarchy of dependencies	Over fitting is a possibility, and the process is computationally intensive.

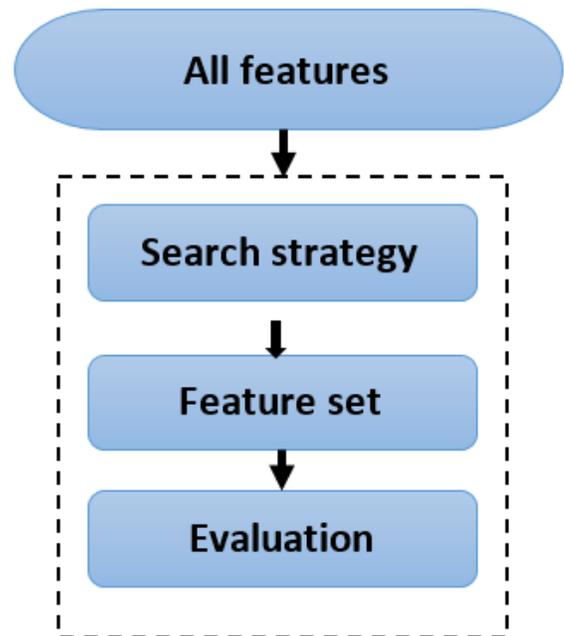


Fig. 1. Filter approach of feature selection

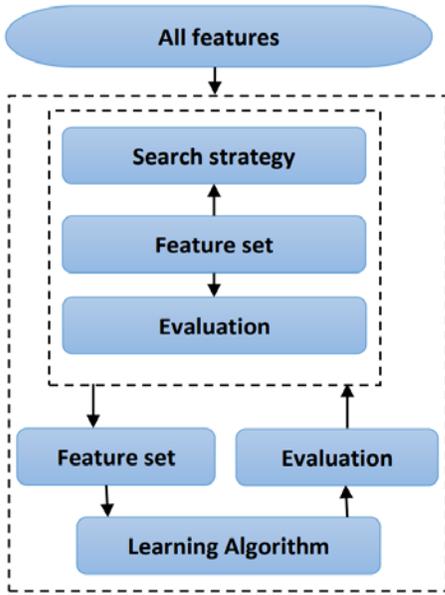


Fig. 2. Wrapper approach of feature selection

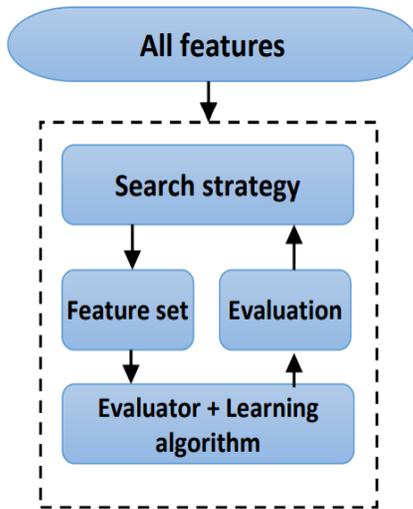


Fig. 3. Embedded approach for feature selection

### B. DBN

Using an unsupervised machine learning paradigm, deep belief networks are a more advanced kind of generative neural network. A DBN can be created by stacking and training individual Restricted Boltzmann Machines (RBM) in a layered configuration. The pre-training stage is a step of unsupervised learning. Stacking sub networks, each with two processing levels, divides the network into groups. These weights are provided to the network as a way to avoid the issues that can arise from using random numbers to initialize the connection weights. Unsupervised learning is used to steer the learning phase of an energy-based stochastic neural network (RBM), which includes two layers of neurons, hidden and visible nodes. Unsupervised training may be used for this function by using the Greedy Layer Wise unsupervised training algorithm. RBM has a hidden layer including nodes and a visible layer including nodes, respectively.

Using an unsupervised learning technique, one can learn an RBM, a generative stochastic neural network. There are two processing tiers in the RBM's network, as depicted in Fig. 4. Construction and reconstruction operations can be carried out independently of one another because these layers are linked together [29].

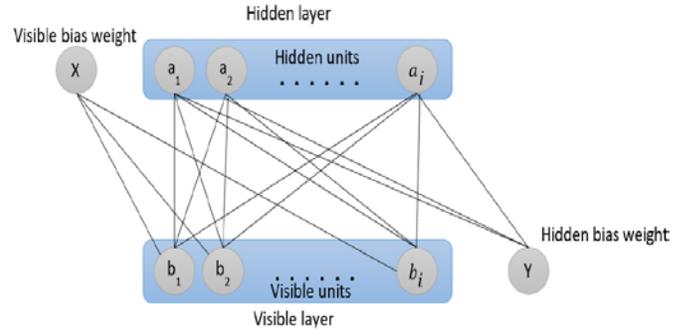


Fig. 4. The RBM architecture

In the visible layer (b), there are visible units  $(b_1, b_2, b_3, \dots, b_i)$  that represent the features of the pattern, while in the hidden layer (a), there are hidden units  $(a_1, a_2, a_3, \dots, a_j)$  that accept their data from the visible units and are able to reconstruct the pattern's characteristics from them. The visible node  $B_i$  and the hidden node  $H_j$  have a weight of  $U_{ij}$  in the  $k \times l$  matrix  $U$ , which represents the weights between visible and hidden levels. Let's assume that the binary and hidden units are  $B$  and  $A$ . In this case,  $A \in \{0,1\}^l$  and  $B \in \{0,1\}^k$ . As shown below in (1), the RBM energy function is a quadratic function of the square of:

$$\text{Ene}(B, A) = -\sum_{i=1}^k R_i B_i - \sum_{j=1}^l Z_j A_j - \sum_{i=1}^k \sum_{j=1}^l B_i A_j U_{ij} \quad (1)$$

where  $Z$  and  $R$  are the basis vectors of the hidden and visible layers, respectively. Also, the likelihood of the  $(B, A)$  configuration is shown in the following equation.

$$\text{pr}(B, A) = \frac{e^{-\text{Ene}(B, A)}}{\sum_{B, A} e^{-\text{Ene}(B, A)}} \quad (2)$$

Normalization is reflected in the denominator of the aforesaid equation. Stochastic gradient descent (SGD) is used to optimize RBM parameters  $Z, R$ , and  $U$  based on the training data's log likelihood. In order to calculate the probability of a given sample over all potential hidden vectors, one can use the following formula:

$$\text{pr}(B) = \frac{\sum_B e^{-\text{Ene}(B, A)}}{\sum_{B, A} e^{-\text{Ene}(B, A)}} \quad (3)$$

$Z, R$ , and  $U$  are used to generate stochastic gradient ascent derivatives of  $\text{pr}(B)$ , which lead to the following equations:

$$U^{n+1} = U^n + \xi(\text{pr}(A/B)B^T - \text{pr}(\tilde{A}/\tilde{B})\tilde{B}^T) - \eta U^T + \alpha \Delta U^{n-1} \quad (4)$$

$$Z^{n+1} = Z^n + \xi(B - \tilde{B}) + \alpha \Delta Z^{n-1} \quad (5)$$

$$R^{n+1} = R^n + \xi(\text{prob}(A/B) - \text{prob}(\tilde{A}/\tilde{B})) + \alpha \Delta R^{n-1} \quad (6)$$

With

$$\text{pr}(A_j = 1/B) = \text{sigma}(\sum_{i=1}^k U_{ij}B_i + R_j) \quad (7)$$

And,

$$\text{pr}(B_j = 1/A) = \text{sigma}(\sum_{i=1}^l U_{ij}A_i + Z_j) \quad (8)$$

where  $\eta$ ,  $n$ ,  $\xi$ , and  $\alpha$  denote the weight decays, number of hidden nodes, learning rate, and momentum weights, in that order. The Softmax function is symbolised by the symbol  $\text{sigma}$  in the logistic model. For the visible and hidden nodes, two different learning methods are used to figure out their weights and biases. Both CD and persistent contrastive divergence (PCD) fall under this category (PCD).

DBN uses an error propagation method and fine-tuned optimal performance to identify the starting weights, which are obtained by undertaking unsupervised pre-training. RBM pre-training is still lacking in the optimal number of layers and nodes. The outcomes are affected by the number of layers and also the number of nodes, but the ideal value relies on the type of dataset and the attributes to be learned. Therefore obtaining the global optimum value has certain downsides. GOA is used to solve the problem of DBN in our job. GOA is a technique for determining the DBN's ideal value and, as a result, reducing error.

#### IV. PROPOSED MODEL

The proposed DFS-ODBN model consists of three phases namely: preprocessing phase, feature selection phase and classification phase. The first phase, preprocessing, involve dealing with the dataset nature which include many missing data due to the archive of data in hospitals. Preprocessing also include normalization of scale of data to make all attribute in a specified range. The second phase involves selecting a subset of attributes to reduce the complexity and time of classification phase. Feature selection phase involve selecting the most appropriate features among all available features using DFS feature selection algorithm in wrapper approach which repeatedly apply DFS method. In last phase, classification, the DBN classifier is used to predict the case of data which ckd or NOT ckd. The classifier parameters are estimated and tuned using an optimization algorithm called GOA. The use of optimization algorithm helps in parameter adaptation to increase the performance of the classifier. The last step is model evaluation which assesses the performance of the proposed system according to many metrics such as: accuracy, sensitivity, f-measure, and precision. The evaluation proved that the proposed system is better than many other related methods. The proposed system is shown in Fig. 5.

##### A. Description of Dataset

UCI's Machine Learning Repository now has the CKD dataset used in this study which was collected and uploaded by the Apollo Hospital in India in 2015. There are 400 data points total, with 25 different properties, 11 of which are numeric and 14 of which are nominal [30]. 250 of the dataset's 400 instances have been assigned to the ckd class, while the remaining 150 have been assigned to the NOTckd class. Table II shows the CSD dataset's attribute breakdown.

##### B. Preprocessing

The quality of the data used in data mining operations must be high in order to achieve a high level of performance at a cheap cost. Anomaly type characteristics will be converted to numeric in the preprocessing step. A total of 14 nominal attributes will be transformed into numerical attributes.

TABLE II. CHRONIC KIDNEY DISEASE DATASET INFORMATION

Features	Type
Red blood cell count (Rc)	Numeric
Sugar (Su)	Nominal
Hypertension (Htn)	Nominal
Sodium (Sod)	Numeric
Red blood cells (Rbc)	Nominal
Packed cell volume (Pcv)	Numeric
Pus cell (Pc)	Nominal
Age (Age)	Numeric
Appetite (appet)	Nominal
White blood cell (Wbcc)	Numeric
Bacteria (Ba)	Nominal
Diabetes mellitus (Dm)	Nominal
Specific gravity (Sg)	Nominal
Serum creatinine (Sc)	Numeric
Anemia (Ane)	Nominal
Pus cell clumps (Pcc)	Nominal
Blood glucoses (Bgr)	Numeric
Coronary artery disease (Cad)	Nominal
Pedal edema (Pe)	Nominal
Blood pressure (Bp)	Numeric
Blood urea (Bu)	Numeric
Albumin (Al)	Nominal
Haemoglobin (Hemo)	Numeric
Potassium (Pot)	Numeric
Class (class)	Nominal

1) *Missing values*: More than half of the variables in the CKD dataset are missing, necessitating the handling of missing values in order to improve accuracy. The mode method is used to replace the empty value with the attribute's maximum frequency when a value is lacking. Attributes can be univariate, monotonous in their missing values, or arbitrary. Only one characteristic has all of the missing values in univariate analysis (feature). If at least three attributes are missing values, the model is said to be monotonous. If the missing values are of random characteristics, then it is arbitrary [31].

2) *Data normalization*: There are numerous approaches to data normalization. Keep the data in a range for each input feature in order to reduce the neural network's preference for one feature over another. Training time can be reduced by normalizing data such that all features are trained at once. It is particularly beneficial for modelling applications when the

inputs are often on a wide range of scales. The features or outputs are rescaled using the Min-Max normalizing method from one range of values to another. Most of the time, the features are rescaled to fall between 0 and 1 or -1 and 1. It is common to perform the rescaling by applying a linear interpretation formula like:

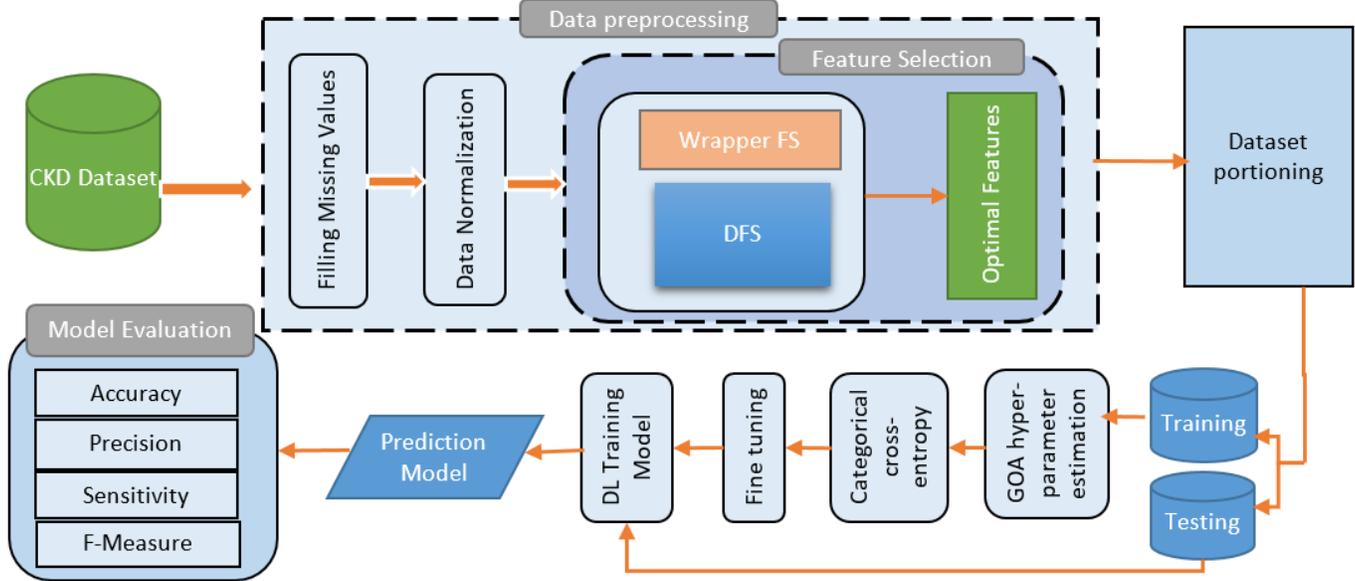


Fig. 5. Proposed framework

$$x'_i = ((\max_{\text{target}} - \min_{\text{target}})x \frac{(x_i - \min_{\text{value}})}{(\max_{\text{value}} - \min_{\text{value}})} + \min_{\text{target}}) \quad (9)$$

where  $(\max_{\text{value}} - \min_{\text{value}}) = 0$  . when  $(\max_{\text{value}} - \min_{\text{value}}) = 0$  for a feature, it shows that that feature in the data has a constant value. Feature values having a constant value should be deleted from the data set because they do not contribute any useful information to the neural network. Min-max normalization maintains the same range of values for each feature when it is applied. The advantage of using min-max normalization is that it keeps all of the data's relationships intact.

### C. DFS for Feature Selection

In order to select the best features, the proposed algorithm employs the following procedures. It is necessary to organize the CKD dataset into groups once the preprocessing is complete. Selecting a collection of features in each iteration is done using DFS. The most important feature in the classification process is a subset of the best features in the raw dataset.

Features can be evaluated using the DFS method, a heuristic approach. A feature is considered good if it has less overlap with other classes than other features. The DFS method takes into account the distribution of features across all classes and their associations for determining rankings. As a first stage in DFS, every feature in every class is given a probability density function (PDF). The next step is to rank the features depending on the overlap area, and this is done next. Parametric and non-parametric methods of computing PDF are

the most frequent approaches [32]. The first method assumes that data has a Gaussian distribution, so the work of estimating density simply entails choosing appropriate values for the distribution's mean and variance. Instead of making assumptions about the shape of the density function, non-parametric techniques simply calculate the density from the observed data. Many pattern recognition applications lack a standard format for estimating the density of raw data. When using random distributions and non-parametric approaches, it is unnecessary to know the fundamental density forms before using these methods [32]. Because of this, the proposed solution is described as using the following parametric approach:

$$p(x) \cong \frac{k}{MVo}$$

Here, the value of the derived PDF for instance  $x$  is represented by the expression  $p(x)$ , whereas  $M$  the total number of examples,  $Vo$  is the volume surrounding  $x$ , and  $k$  the number of instances within  $V$  are all given. To get a more accurate PDF, try increasing  $M$  and decreasing  $Vo$ . After estimating PDFs for each class, the next step is to compare the value of each feature based on the calculated PDFs for each class. As previously stated, a feature is considered good if it has less class overlap than the rest. The estimation of PDF for each class label and feature is used to estimate the amount of overlap between occurrences of a certain feature class. The significance of a feature for class label prediction decreases as the overlapping region grows larger, and this leads to a decline in classification performance. The overlapping value of a feature  $r$  in class  $\omega$  can be calculated using the formula 2.

$$Ov(r, \omega) = \int \text{Minimum}(\text{Maximum}(\text{PDF}(\omega_i)), \text{PDF}(\omega))$$

such that  $i$  is not equal to  $\omega$  and  $1 \leq i \leq \#$  of classes

#### D. Optimized DBN based on GOA

"Reference [33] proposed the GOA algorithm. It was inspired by the swarming activity of grasshoppers in nature. A grasshopper's flight path in a swarm is affected by the following three factors: Those three factors are: human social connection, gravitational pull, and wind advective forces. The GOA algorithm is used for minimizing the value of error in order to get the ideal DBN value. The GOA has a series of steps, as outlined below:

1) *Step 1: initial step:* It is necessary to initialize the RBM parameters in GOA, as well as the number of candidate solutions,  $MIN_{CA}, MAX_{CA}$ , and maximum number of evaluations, before performing any analysis. A decrease in the repulsion area, comfort area, and attraction area is represented here by the parameter  $CA$ . The starting population is then generated at random, and using the objective function, each solution in the population is assessed.

2) *Step 2: assessment of fitness:* For each search agent, the fitness function is calculated and computed after initialization. For the sake of this study, we define fitness as the minimization of DBN's mean square error:

$$\text{Fitness} = \text{Min}_{\text{error}} = \frac{1}{E} \sum_{i=1}^E \left( \frac{P(x,y) - T_y}{T_y} \right)^2 \quad (10)$$

The mean square error denotes the average distance between the predicted and observed values.  $T_y$  and  $P(x, y)$  represent the target value for the appropriate data  $y$  and the estimated probability of the appropriate data  $x$ , respectively.

The RBM parameter is tuned using this algorithm to reduce the amount of error that occurs. Using a DBN for unsupervised pertaining and supervised fine tuning was the basis for the proposed method. GOA is used throughout the entire process of relating and fine tuning. The range of hyper-parameters in our model was identified by conducting random search experiments and then selecting their values at random until we achieved the best performance. It takes less time and effort to train a network using random search. Because not all hyper-parameters are equally important for tuning, random experimentation based on hyper-parameter values is more efficient.

3) *Step 3: updating:* Update the best target's position in each evaluation and factor  $C$  in this phase. In actuality, there is no goal because we don't know what the global optimum is or what the precise goal is. The target must be identified at each stage of the optimization process. During optimization in GOA, it is assumed that the fittest grasshopper is the target. Grasshopper must advance toward the good target in order for GOA to save it in search space during this iteration. Position and  $C$  updates are calculated as follows:

$$x_m^d = C \left( \sum_{n=1, n \neq m}^Z C \frac{d_{ub} - d_{lb}}{S} S(|X_n^d - X_m^d|) \frac{X_n - X_m}{d_{mn}} \right) + d_T \quad (11)$$

In the  $d$ -dimension,  $d_{lb}$  and  $d_{ub}$  denote the lower and upper bounds, respectively. The  $d$ -dimensional goal is  $d_T$ . The following formula can be used to determine the strength of social forces represented by the function  $S$ :

$$S(f) = g e^{-\frac{f}{r}} - e^{-f} \quad (12)$$

where  $g$  and  $r$  stand for the intensity of attraction and the length scale of attraction, respectively.  $C$  represents a decreasing coefficient that reduces the repulsion zone, comfort zone, and attraction zone. The GOA algorithm's primary governing parameter is  $C$ , which is kept current using the eq. 13 below.

$$C = \text{Max}_C - y \frac{\text{Max}_C - \text{Min}_C}{Y} \quad (13)$$

where  $Y$  denotes the most iterations possible and the current iteration is referred by  $y$ ,  $\text{Max}_C = 1$  and  $\text{Min}_C = 0.00001$ .

4) *Step 4: termination:* When the maximum number of iterations is achieved, the position is updated iteratively. Finally, the best objective was returned, with the global optimum being a combination of position and fitness. The GOA algorithm is used to figure out what the ideal value of DBN should be. Once DBN has classified the ckd or NOTckd, the process is complete. This means that the result from the classifier is either positive or negative. Below, the GOA algorithm 1 is illustrated and the steps are shown in "Fig. 6".

Line #	Algorithm 1: DBN optimized based on GOA algorithm
1	<b>Input:</b> candidate solution, $RBM$ parameters, $MAX_C$ , maximum # of iters, $MIN_C$ ,
2	<b>Output:</b> Optimal parameter combination
3	Set initialization of the candidate solution and random population of parameters of RBM
4	Set initialization of $MAX_C$ , maximum # of iters, $MIN_C$
5	Compute the fitness value $fit(i)$ for each individual agent using mean square error as fitness function
6	Set best individual search agent = $W$
7	<b>While loop</b> ( $Q < \text{max \# of iters}$ )
8	modify $C$ using eq. 3
9	<b>foreach</b> individual:
10	The current individual position is modified using eq. 2
11	reset the individuals above and below the bounds
12	<b>for end</b>
13	Updating $W$ to a better solution
14	$Q = Q + 1$
15	<b>while end</b>
16	Return $W$ as optimal

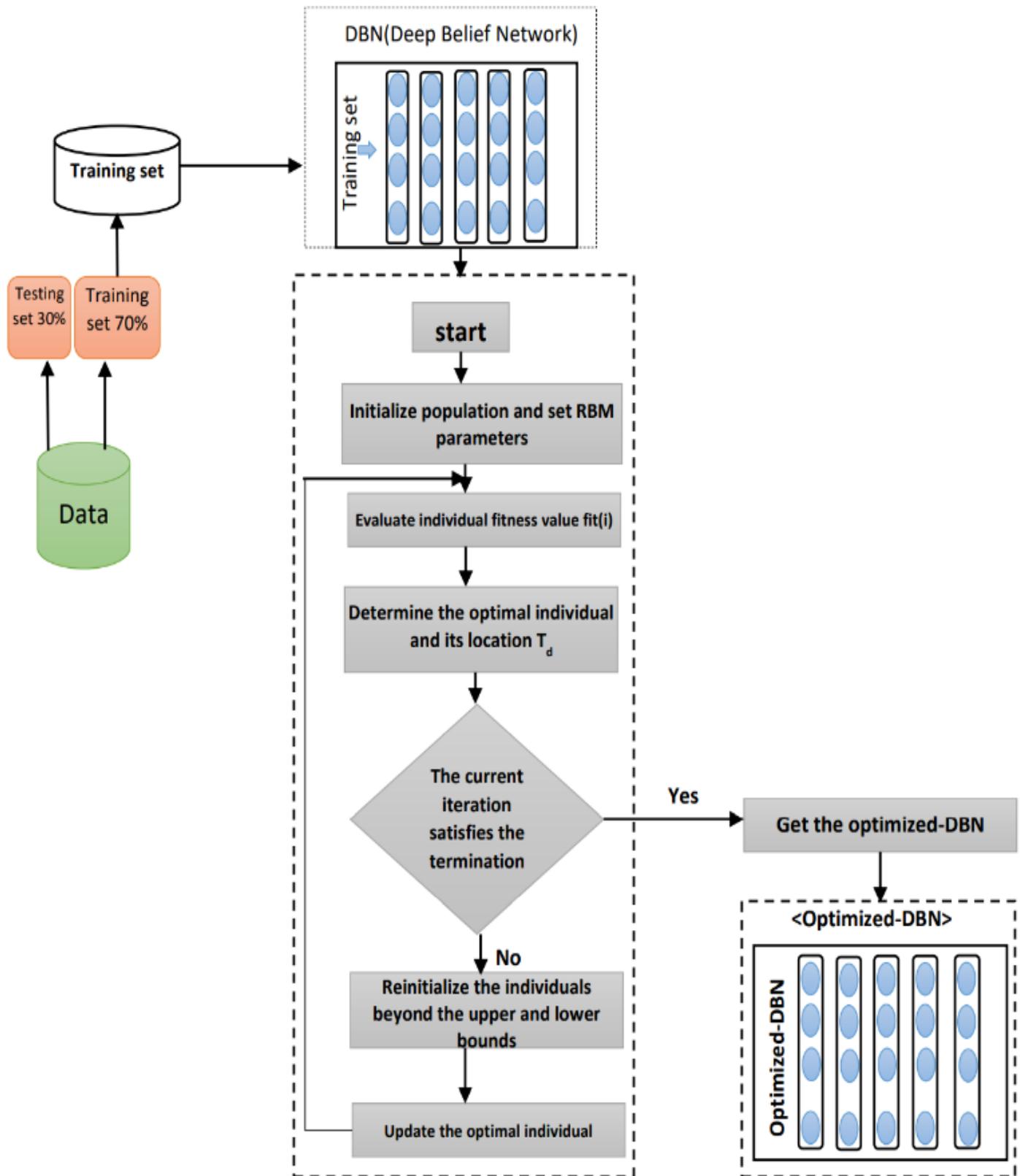


Fig. 6. DBN optimized based on GOA algorithm

## V. EXPERIMENTAL RESULTS

### A. Environmental Setup

The proposed model was developed in MATLAB 2016a, with some preprocessing done in Weka. MATLAB is a powerful data mining tool. Models and applications are built based on data analysis. A Core i7 machine with an NVIDIA graphics card is utilized for testing and assessments.

### B. Performance Metrics

In order to train the model, the data set is randomly divided into two parts, with the first portion containing 70% of all the total collection of data. The data from the second section is used for testing (30% of the time). The suggested model is evaluated and validated using six performance measures. Accuracy, precision, sensitivity, and F-Measure [34] are some examples of these parameters. By measuring performance measures, the confusion matrix describes the performance of categorization algorithms. The following measures were utilized in this study to evaluate the performance of the methodologies in use.

1) *True positive (TP)*: This indicates occurrences of positive outputs that have been appropriately classified.

2) *False negatives (FN)*: These are false negatives that are not actually false negatives.

When unfavorable outcomes are mistakenly labeled as positive ones, the term "False Positive" (FP) is used.

3) *False negative (FN)*: good events that were mistakenly labeled as negative in the report.

- **Accuracy**: An image's accuracy in a database is determined by how closely the image's coordinates match the database's real value. Accuracy measures are quite close to the genuine value, and they are processed as a true proportion of the outcomes:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (14)$$

- **Sensitivity (Recall)**: Sensitivity is a term used to describe the state of being sensitive. A test's sensitivity or recall refers to its capacity to correctly identify people who suffer from a certain illness (True Positive Rate). In this way, it can be stated:

$$\text{Sensitivity} = \frac{TP}{TP+FN} \quad (15)$$

- **Specificity**: When a test has high specificity, it can accurately identify people who do not have the condition (True Negative Rate). According to the definition, it is:

$$\text{Specificity} = \frac{TN}{TN+FP} \quad (16)$$

- **Precision**: Predictive Value (PPV) or Precision is a measure of the accuracy of a categorization result. The following is the formula used to arrive at this result:

$$\text{Precision} = \frac{TP}{TP+FP} \quad (17)$$

- **F-measure**: According to this method, the harmonic mean of precision and recall is calculated as follow:

$$F - \text{Measure} = 2 \times \frac{\text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}} \quad (18)$$

### C. Performance Comparison

As a result of using a variety of feature selection methods, including filtering and wrapping, CKD diagnosis was improved. In all of the methods used, the original dataset can be reduced in dimension to produce a new dataset. Twenty-five different variables were included in the dataset that was used. The dataset was condensed down to 11 attributes using the DFS wrapper approach. On a smaller dataset, the optimized-DBN classifier was able to identify 2 FPs and 3 FNs. The value of FN in the Optimized-DBN classifier method is lower than the values of FN in the other methods. When compared to the current system, the proposed combination of Optimized-DBN and unsupervised training has better accuracy than other research work. Layers (visible layer, hidden layer, and output layer), nodes, and weights and biases for the layers were used to develop the final architecture of Optimized-DBN. Our competitors' quantitative measures are compared to ours in this section. Table III represents the different performance evaluation metrics, i.e., accuracy, f-measure, precision, and sensitivity for the proposed system DFS-ODBN and other different related methods.

A sensitivity and accuracy analysis of the proposed DFS-ODBN technique is shown in Fig. 7. As shown in the figure, the FDS-ODBN technique has improved in both sensitivity and precision.

TABLE III. PERFORMANCE METRICS FOR PROPOSED SYSTEM AND OTHER SYSTEMS

Methods	Results(%)			
	Sensitivity	Precision	F-measure	Accuracy
Shrivastava, Sahu and Hota [11]	94.00	95.14	94.57	93.25
Rady and Anwar [14]	95.61	95.98	95.79	94.75
Rubini and Perumal [16]	91.61	92.33	91.97	95.00
Elhoseny et al [18]	96.00	96.46	96.00	95.00
Chittora et al [20]	98.00	96.67	98.31	98.86
M.M. Hossain [22]	99.43	99.82	99.60	99.50
Proposed model	99.63	98.81	99.63	99.70

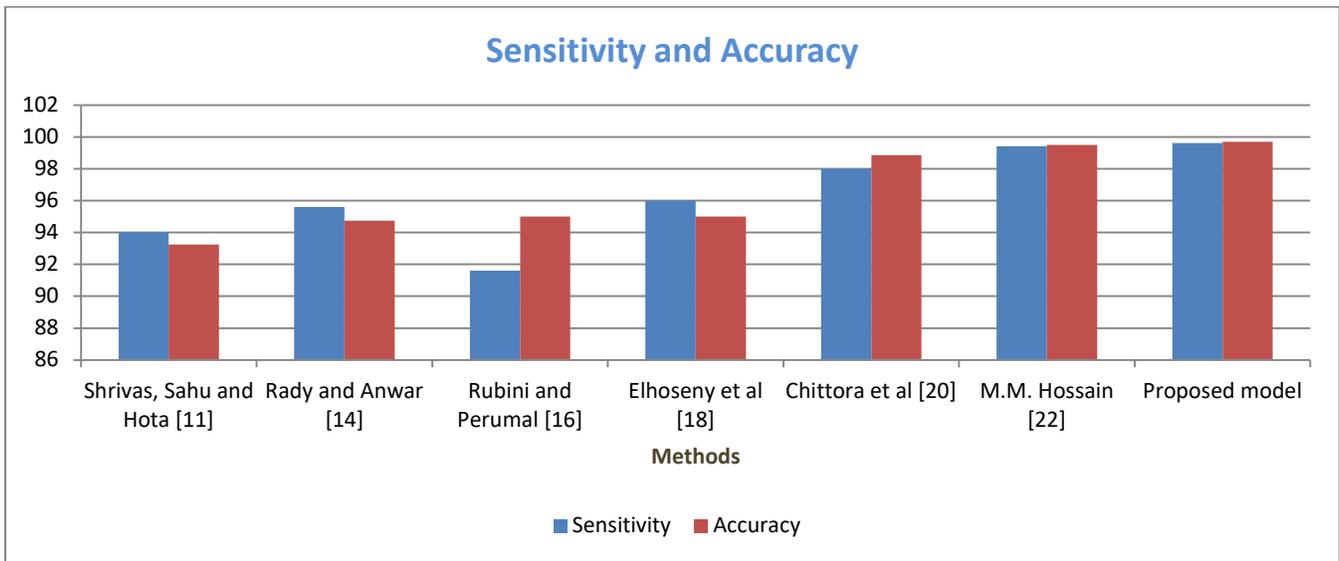


Fig. 7. Evaluation metrics (sensitivity and accuracy) Vs. classifiers

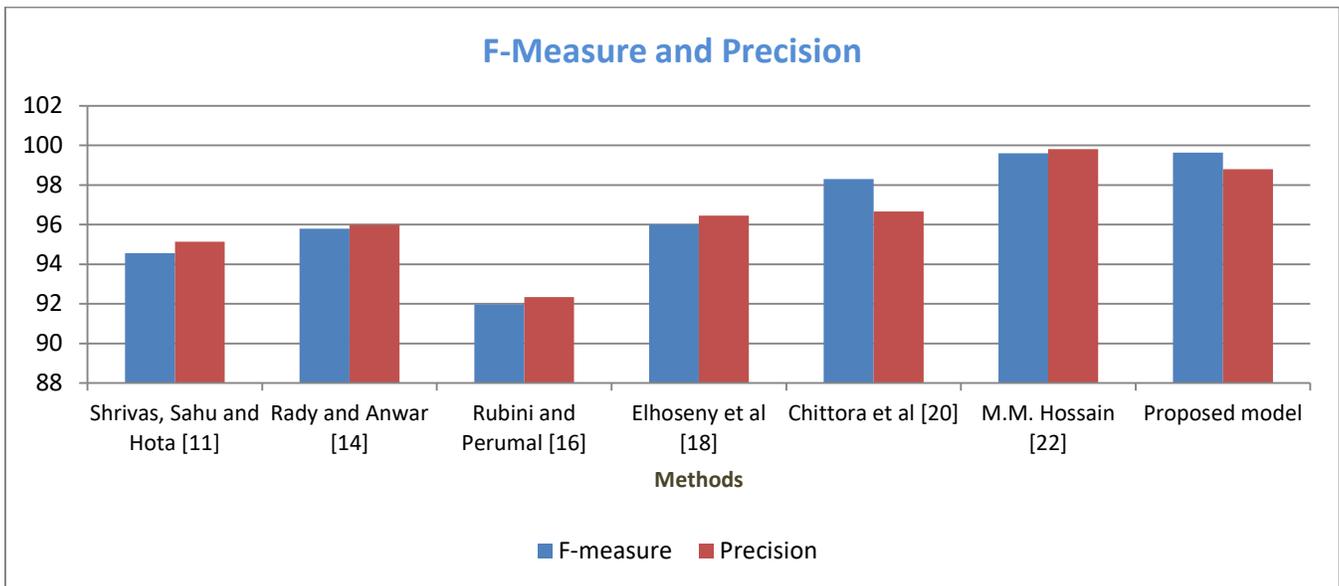


Fig. 8. Evaluation metrics (F-measure and precision) Vs. classifiers

This method is, 4.7 percent more accurate than [16]. 6.52 percent more accurate than [11]. percent more accurate than [14]. 4.7 percent higher [18] and 1.1 percent higher [20].

F-measure and precision analysis are shown in Fig. 8 in terms of the feature selected for the proposed technique DFS-ODBN. According to the graph, the FDS-ODBN technique now has higher precision and precision values. In terms of precision, this method is 3.66308 percent higher than [11], and 2.82084 percent higher than [14]. The proposed method is 6.46607 percent higher [16] and 3.364 percent higher [18] in terms of the F-measure.

## VI. CONCLUSION AND FUTURE WORK

The study looked at the classification of medical data in order to identify the patient's disease at an early stage of development. The most difficult part of classifying medical

data is selecting the best subset of attributes from the dataset under consideration. A DFS feature selection algorithm was used to select the best features from a preprocessing stage in which the missing values were eliminated. Based on the presence or absence of CKD, the dataset was divided into two classes: the ckd class and the NOT ckd class. The Deep Belief Network algorithm was used for this classification because it is the best method for data classification. It was necessary to use the GOA algorithm in order to obtain the DBN network hyperparameters. GOA has strong capabilities to explore the search space and it benefits from high exploration and exploitation.

Furthermore, the classification issues can be overcome with an average computational cost. Using CKD datasets, Optimized-DBN was able to achieve its maximum performance in terms of sensitivity, accuracy, and specificity. When compared to other techniques, the proposed DFS-ODBN

demonstrated accuracy of 99.75 percent. In the future, with algorithms designed or prediction techniques, the data classification can be enlarged and missing values can be removed by new imputation approaches, classification, and prediction. The development of hybrid and novel optimization algorithms for the classification of medical data and feature selection is recommended as a focus for future contributions

#### REFERENCES

- [1] H. Polat, H. D. Mehr, and A. Cetin, "Diagnosis of chronic kidney disease based on support vector machine by feature selection methods," *J. Med. Syst.*, vol. 41, no. 4, p. 55, 2017.
- [2] A. S. Go, G. M. Chertow, D. Fan, C. E. McCulloch, and C. Hsu, "Chronic kidney disease and the risks of death, cardiovascular events, and hospitalization," *N. Engl. J. Med.*, vol. 351, no. 13, pp. 1296–1305, 2004.
- [3] P. Kathuria and B. Wedro, "Chronic kidney disease quick overview." IOP Publishing [edicinehealth](https://doi.org/10.1063/1.5000000), 2016.
- [4] M. H. A. Elhebir and A. Abraham, "A novel ensemble approach to enhance the performance of web server logs classification," *Int. J. Comput. Inf. Syst. Ind. Manag. Appl.*, vol. 7, no. 1, pp. 189–195, 2015.
- [5] G. A. Afzali and S. Mohammadi, "Privacy preserving big data mining: association rule hiding using fuzzy logic approach," *IET Inf. Secur.*, vol. 12, no. 1, pp. 15–24, 2018.
- [6] M. Shardlow, "An analysis of feature selection techniques," *Univ. Manchester*, vol. 1, no. 2016, pp. 1–7, 2016.
- [7] I. Guyon, S. Gunn, M. Nikravesh, and L. A. Zadeh, *Feature extraction: foundations and applications*, vol. 207. Springer, 2008.
- [8] R. Jensen, "Combining rough and fuzzy sets for feature selection," *Citeseer*, 2005.
- [9] B. Xue, M. Zhang, W. N. Browne, and X. Yao, "A survey on evolutionary computation approaches to feature selection," *IEEE Trans. Evol. Comput.*, vol. 20, no. 4, pp. 606–626, 2015.
- [10] D. Mladenović, "Feature selection for dimensionality reduction," in *International Statistical and Optimization Perspectives Workshop Subspace, Latent Structure and Feature Selection*, 2005, pp. 84–102.
- [11] A. K. Shrivastava, S. K. Sahu, and H. S. Hota, "Classification of chronic kidney disease with proposed union based feature selection technique," in *Proceedings of 3rd International Conference on Internet of Things and Connected Technologies (ICIoTCT)*, 2018, pp. 26–27.
- [12] C. Xiong, M. Su, Z. Jiang, and W. Jiang, "Prediction of hemodialysis timing based on LVW feature selection and ensemble learning," *J. Med. Syst.*, vol. 43, no. 1, pp. 1–8, 2019.
- [13] S. Ravizza et al., "Predicting the early risk of chronic kidney disease in patients with diabetes using real-world data," *Nat. Med.*, vol. 25, no. 1, pp. 57–59, 2019.
- [14] E.-H. A. Rady and A. S. Anwar, "Prediction of kidney disease stages using data mining algorithms," *Informatics Med. Unlocked*, vol. 15, p. 100178, 2019.
- [15] G. R. Vásquez-Morales, S. M. Martínez-Monterrubio, P. Moreno-Ger, and J. A. Recio-García, "Explainable prediction of chronic renal disease in the colombian population using neural networks and case-based reasoning," *IEEE Access*, vol. 7, pp. 152900–152910, 2019.
- [16] L. J. Rubini and E. Perumal, "Hybrid kernel support vector machine classifier and grey wolf optimization algorithm based intelligent classification algorithm for chronic kidney disease," *J. Med. Imaging Heal. Informatics*, vol. 10, no. 10, pp. 2297–2307, 2020.
- [17] L. Jerlin Rubini and E. Perumal, "Efficient classification of chronic kidney disease by using multi-kernel support vector machine and fruit fly optimization algorithm," *Int. J. Imaging Syst. Technol.*, vol. 30, no. 3, pp. 660–673, 2020.
- [18] M. Elhoseny, K. Shankar, and J. Uthayakumar, "Intelligent diagnostic prediction and classification system for chronic kidney disease," *Sci. Rep.*, vol. 9, no. 1, pp. 1–14, 2019.
- [19] B. Khan, "Empirical Evaluation of ML Techniques for CKD Prophecy," *IEEE Access*, vol. 8, March 2020, DOI:10.1109/ACCESS.2020.2981689.
- [20] P. Chittora et al., "Prediction of Chronic Kidney Disease - A Machine Learning Perspective," *IEEE Access*, vol. 9, January 2021, DOI 10.1109/ACCESS.2021.3053763.
- [21] E. M. Senan, M. H. Al-Adhaileh, and F. W. Alsaade, "Diagnosis of Chronic Kidney Disease Using Effective Classification Algorithms and Recursive Feature Elimination Techniques," *Journal of Healthcare Engineering*, vol. 2021, <https://doi.org/10.1155/2021/1004767>.
- [22] M. M. Hossain, Reshma. A. Swarna, "Analysis of the performance of feature optimization techniques for the diagnosis of machine learning-based chronic kidney disease," *Machine Learning with Applications*, <https://doi.org/10.1016/j.mlwa.2022.100330>.
- [23] R. K. Singh and M. Sivabalakrishnan, "Feature selection of gene expression data for cancer classification: a review," *Procedia Comput. Sci.*, vol. 50, pp. 52–57, 2015.
- [24] C.-T. Su and C.-H. Yang, "Feature selection for the SVM: An application to hypertension diagnosis," *Expert Syst. Appl.*, vol. 34, no. 1, pp. 754–763, 2008.
- [25] B. Kumari and T. Swarnkar, "Filter versus wrapper feature subset selection in large dimensionality micro array: A review," 2011.
- [26] O. Villacampa, "(Weka - Thesis) Feature Selection and Classification Methods for Decision Making: A Comparative Analysis," *ProQuest Diss. Theses*, no. 63, p. 188, 2015, [Online]. Available: [https://login.pallas2.tcl.sc.edu/login?url=https://search.proquest.com/docview/1721469983?accountid=13965%0Ahttp://resolver.ebscohost.com/olpurl?ctx\\_ver=Z39.88-2004&ctx\\_enc=info:ofi/enc:UTF8&rft\\_id=info:sid/ProQuest+Dissertations+%26+Theses+Global&rft\\_v](https://login.pallas2.tcl.sc.edu/login?url=https://search.proquest.com/docview/1721469983?accountid=13965%0Ahttp://resolver.ebscohost.com/olpurl?ctx_ver=Z39.88-2004&ctx_enc=info:ofi/enc:UTF8&rft_id=info:sid/ProQuest+Dissertations+%26+Theses+Global&rft_v).
- [27] A. G. Karegowda, A. S. Manjunath, and M. A. Jayaram, "Feature Subset Selection Problem using Wrapper Approach in Supervised Learning," *Int. J. Comput. Appl.*, vol. 1, no. 7, pp. 13–17, 2010, doi: 10.5120/169-295.
- [28] S. Khalid, T. Khalil, and S. Nasreen, "A survey of feature selection and feature extraction techniques in machine learning," *Proc. 2014 Sci. Inf. Conf. SAI 2014*, no. October, pp. 372–378, 2014, doi: 10.1109/SAI.2014.6918213.
- [29] G. E. Hinton, "A practical guide to training restricted Boltzmann machines," in *Neural networks: Tricks of the trade*, Springer, 2012, pp. 599–619.
- [30] J. Li et al., "Imputation of missing values for electronic health record laboratory data," *NPJ Digit. Med.*, vol. 4, no. 1, pp. 1–14, 2021.
- [31] T. Jo and N. Japkowicz, "Class imbalances versus small disjuncts," *ACM Sigkdd Explor. Newsl.*, vol. 6, no. 1, pp. 40–49, 2004.
- [32] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification. 605 Third Avenue.* New York, USA: John Wiley & Sons, 2012.
- [33] S. Saremi, S. Mirjalili, and A. Lewis, "Grasshopper optimisation algorithm: theory and application," *Adv. Eng. Softw.*, vol. 105, pp. 30–47, 2017.
- [34] M. Liu, C. Xu, Y. Luo, C. Xu, Y. Wen, and D. Tao, "Cost-sensitive feature selection via f-measure optimization reduction," 2017.

# Automated Categorization of Research Papers with MONO Supervised Term Weighting in RECApp

Ivic Jan A. Biol<sup>1</sup>, Rhey Marc A. Depositario<sup>2</sup>, Glenn Geo T. Noangay<sup>3</sup>, Julian Michael F. Melchor<sup>4</sup>,  
Cristopher C. Abalorio<sup>5</sup>, James Cloyd M. Bustillo<sup>6</sup>

Computer Science Program, Computer Education Department,  
ACLC College of Butuan, Butuan City, Philippines<sup>1, 2, 3, 4, 5, 6</sup>  
Caraga State University, Butuan City, Philippines<sup>5</sup>  
Graduate Programs, Technological Institute of the Philippines, Quezon City<sup>5, 6</sup>

**Abstract**—Natural Language Processing, specifically text classification or text categorization, has become a trend in computer science. Commonly, text classification is used to categorize large amounts of data to allocate less time to retrieve information. Students, as well as research advisers and panelists, take extra effort and time in classifying research documents. To solve this problem, the researchers used state-of-the-art supervised term weighting schemes, namely: TF-MONO and SQRTF-MONO and its application to machine learning algorithms: K-Nearest Neighbor, Linear Support Vector, Naive Bayes Classifiers, creating a total of six classifier models to ascertain which of them performs optimally in classifying research documents while utilizing Optical Character Recognition for text extraction. The results showed that among all classification models trained, SQRTF-MONO and Linear SVC outperformed all other models with an F1 score of 0.94 both in the abstract and the background of the study datasets. In conclusion, the developed classification model and application prototype can be a tool to help researchers, advisers, and panelists to lessen the time spent in classifying research documents.

**Keywords**—Text classification; supervised term weighting schemes; optical character recognition; machine learning algorithms

## I. INTRODUCTION

Writing and publishing have taken popularity on the internet using online services where text classification plays an important role [1]. An example where text classification can be applied is in the increasing amount of published research documents online or offline due to the advancement of computer and information technologies [2]. Documents, in this case, refer to textual records, and each copy contains a group of words that ranges from sentence to paragraph long. It is through the use of text classification, the prediction and classification of documents can be made possible by categorizing them into which class they belong based on their inherent properties [3].

While there are many ways to classify research papers online, there is also a need to categorize those with only physical copies. Approved research papers refer to peer-reviewed and panel-evaluated complete research in local school libraries. Additionally, research papers still in the proposal period are subject to revisions and need to be more

easily distinguishable whether they are suitable for the course. Moreover, classifying large documents is time and energy-consuming [4]. For such reasons, it is necessary to make a tool that efficiently organizes approved and work-in-progress research papers individually or in bulk.

In this study, to classify research documents, first, OCR will be used. Optical Character Recognition (OCR) acquires an image through the use of a device, usually a camera or scanner, and then converts it to digital text [5][6][7]. Then, supervised term weighting schemes are applied to assign a weight for each term in every document, enhancing text classification performance [8]. The documents will be assigned to their designated classes using different machine learning algorithms [9].

The researchers came up with the idea to design a classification model using different combinations of supervised Term Weighting Schemes (TWS) and Machine Learning Algorithms, as well as prove which combination of Supervised TWS and Machine Learning algorithms is the fastest and yields the results with the highest accuracy for classifying our dataset. Finally, this study will develop an application that will use the designed classifier model to categorize research papers while allowing users to import text images or use the device's camera for text image acquisition.

The subsequent sections of this paper cover various important aspects of the study, including: a literature review on text classification, OCR, term weighting schemes, and machine learning algorithms in Section II, an in-depth examination of the research methods used in the study in Section III, a presentation of the results and discussion of the findings in Section IV, and a conclusion summarizing the key takeaways and recommendations for future research in Section V.

## II. LITERATURE REVIEW

In this section, we present a comprehensive review of relevant literature pertaining to this study. This includes a detailed examination of research on Text Classification, Optical Character Recognition (OCR), Weighting Schemes, and various machine learning algorithms. The sub-sections will thoroughly understand the field's current state and serve as a foundation for this research.

### A. Text Classification

Text classification or also known as text categorization, is a technology for information organization and management wherein it has been proven to be effective and efficient [10]. In natural language processing, text classification is a crucial task and has been its foundation. Over the years, numerous research in this area has been published due to the unprecedented success of deep learning [11].

Through the use of the term frequency-inverse document frequency (TF-IDF), Latent Dirichlet Allocation (LDA), and K-Means clustering, [2] designed a research paper classification model wherein it classifies and clusters similar papers based on its abstract. First, a keyword dictionary which contains groups of keywords that have similarity in meaning, is constructed into one representative keyword or topic. However, yielding results will still cause high running time. In order to solve this problem, they used LDA to extract topic sets before calculating the word and document frequency using TF-IDF. Then, the set of data is classified into classes based on their similarity using K-means clustering, a clustering technique used to minimize distances between every data point and the nearest cluster or centroid. Finally, to evaluate the accuracy of their classification system, they used F-score, an evaluation metric used in text classification, combining precision and recall values.

### B. Optical Character Recognition

Over the past few decades, the area of pattern recognition has been a topic of study which is known as Optical Character Recognition. For many diverse styles of programs in different fields, Optical character recognition is the bottom of it, which we use in our daily life. These days, Optical Character Recognition is being used in many different areas of research [12][13].

P. Divya et al. (2021) developed a web-based optical character recognition application using flask and tesseract [14]. Their website allows users to upload image format data and convert it into machine-editable text in a fraction of a second. It can also read and convert handwritten data with a slightly lower accuracy compared to digitized text. Their OCR model has been proven to be accurate by testing a combined handwritten and digitized test set of 1000 images. In their study, they found that the best input that their system can convert to text is digitized black and white with an accuracy rate of 98%.

### C. Supervised Term Weighting Scheme

The selection of an appropriate term weighting scheme is important in text classification tasks as it has significant effects on its performance. Term Weighting Schemes (TWS) determine how texts would be represented in the vector space model. The terminology Supervised Term Weighting Schemes have been gaining popularity in the recent years while term frequency-inverse document frequency (TF-IDF) is still widely used. But its disadvantage is that it does not train text using the available categories [15][16] and is considered an unsupervised term weighting scheme, unlike the state-of-the-art supervised term weighting schemes: TF-IGM and TF-MONO.

Most of the popular novel supervised TWS focuses on assigning weights based on how the terms occur throughout the classes, which is a proven and effective method in term weighting. However, Dogan and Uysal (2020) believed that this information may not be enough to determine the terms' power in the document [17]. They proposed the novel STW scheme Term Frequency Max-Occurrence and Non-occurrence (TF MONO), which makes use of the non-occurrence information along with the max-occurrence information of terms in the document.

TF-MONO is a supervised TWS that uses the class with the max-occurrence as well as the non-occurrence information in the document frequency. The procedure of the MONO TWS is represented in Fig. 2. Fig. 1 illustrates the MONO TWS designed by [17] separated into seven (7) steps, and visualized by [18][19], which will be further explained in details.

The steps for performing MONO on a text collection are as follows:

- 1) Sort the document frequency of a term in descending order.
- 2) Divide the sorted document frequency into two groups: one for the highest class document frequency values and the other for the rest of the classes.
- 3) Represent the first group with a max-occurrence (MO) ratio and the second group with a non-occurrence (NO) ratio.
- 4) Calculate the MO ratio as the ratio between the quantity of text documents in the class where the term occurs most and its total quantity of text documents.
- 5) Calculate the NO ratio as the ratio between the quantity of text documents in the rest of the classes where the term does not occur and the total quantity of text documents in the rest of the classes.
- 6) Calculate the product of MO and NO ratios and assign it as the MONO (Local) weight of the term.
- 7) Calculate the MONO (Global) weight of the term by using the MONO (Local) weight and a balance parameter  $\alpha$  with a default value of 7.0.
- 8) Finally, two (2) term weighting schemes based on MONO (Global) collection frequency factor are shown.

### D. Machine Learning Algorithms

As the internet expands, the number of unorganized data is also increasing. Thus, intelligent programs that use machine learning in classifying documents have been researched and developed to efficiently access information. Some of the machine learning techniques used for document classification include Naïve Bayes, Support Vector Machine, Decision Trees, etc. [20]. A. Barua et al. (2021) used the most common machine learning methods, namely: Logistic Regression, Support Vector Classifier, Decision Tree(C4.5) [24], Naïve Bayes, Random Forest, and K-Nearest Neighbor, to classify articles about sports into four (4) different categories: Cricket, Football, Tennis, and Athletics. 80% of the data were used for training, while the other 20% were for testing [21]. Using F1-score for evaluation, the result shows that the "Cricket" category has the highest F1 value as it has the most data used

for training among all the classes. Meanwhile, the category with the lowest F1 value is “Athletics” due to its low number of training data. Regarding the machine learning models, Naïve Bayes has the best performance (98.53%) for identifying documents in the cricket class with unigram + bigram + trigram feature while KNN has shown poor performance on an imbalance dataset.

### III. METHODOLOGY

This section outlines the key components integral to the research study, which employs machine learning models for data analysis. These include the data sources, research design, data collection and preprocessing techniques, machine learning models, evaluation metrics, and the application simulation used to test the models. All these components work together to provide a comprehensive framework for conducting a thorough and systematic study, utilizing machine learning models to extract insights from data, and evaluate the performance of the models. The research framework and its implementation are illustrated in Fig. 1, providing a clear understanding of the methodology adopted in the study.

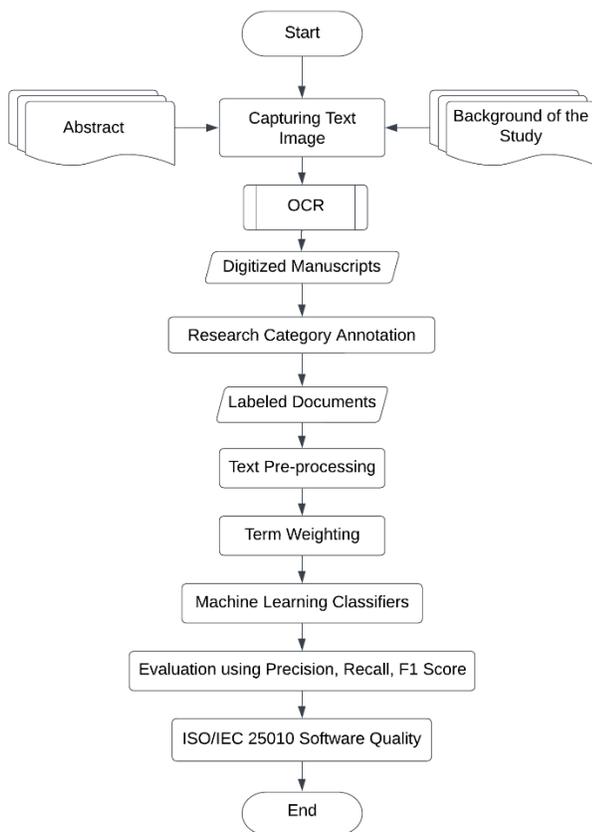


Fig. 1. Framework of the study

The *components* in the framework of the study includes the following:

#### A. Data

The data used in this study is gathered from the Abstract and Background of the Study (BOS) of research documents from ACLC College of Butuan, Saint Joseph Institute of Technology, and Caraga State University. A total of 462

research documents were gathered from the mentioned local schools. To address the lack of data, the researchers used online websites as a source of additional data, with a total of 519 research documents. However, the dataset is still imbalanced, with the feasibility study having only 99 documents. In order to make the dataset unbiased, oversampling is performed on the feasibility study, adding 99 more documents. Overall, there are a total of 1,121 research documents with four (4) categories: Capstone (313), Thesis (306), Case Study (304), and Feasibility Study (198).

#### B. Converting Research Papers to Image Format

Research papers acquired online are in .pdf image format. However, numerous research papers from local schools are not digitized. Therefore, it is necessary to capture images of the research documents' title, abstract, and background of the study, the parts of a research document necessary to be used as our data. The image format of the said images is .jpeg and can still be converted to digital text through OCR.

#### C. PyTesseract OCR

The images were processed one by one through a python script that utilizes PyTesseract, a well-known python Optical Character Recognition library, to convert text images into a digital format. This library allows for accurate and efficient conversion of images to text, making it an ideal choice for this task. The script iterates through all the images, performs the conversion and saves the digital text in a file, ready for further analysis.

#### D. Labeling Research Categories

The converted texts were then placed in a csv file for easy access and organization. The Pandas python library was used to facilitate this process, as it allows for efficient manipulation and storage of large datasets. Each text was labeled manually by researchers according to its predefined category to aid in the analysis and classification process.

#### E. Pre-Processing

Text pre-processing is an integral part of text classification as it can improve the overall quality of a dataset. It can clean the dataset by removing unneeded parts of the text, such as repetitions and spelling errors [22]. Text pre-processing includes four (4) basic processes: tokenization, stop words removal, stemming, and vector space model [23]. Tokenization involves the removal of spaces and taking unique words from the document. Stop word removal is the process of removing stop words which are prevalent words with little to no meaning in the document. Stemming converts all of the words in the document into their root words to reduce the unique words in the document. Vector space modeling, also known as vectorization or term weighting, assigns weight to unique words. To prepare the dataset for this study, the researchers preprocessed the text in the following order:

1) Removing all punctuations and transforming all characters to lowercase to standardize the text and make it easier to work with.

2) Tokenization was applied to split the text into multiple words which makes it more manageable for analysis.

3) Stop words were removed as they do not carry any significant meaning for the analysis.

4) Filtering the features by length was done to remove random words that hold no meaning, this helps to reduce the noise in the dataset.

5) Stemming was applied using Porterstemmer which reduces words to their base form, this helps to group similar words together, and enables better analysis.

These steps helped to clean the dataset, making it more organized and ready for the analysis. The preprocessing steps not only standardize the text but also make it more manageable and focused for analysis, which ultimately results in more accurate and meaningful results.

#### F. Term Frequency Distribution

In order to apply any term weighting scheme, it is crucial to first calculate the term frequency (TF) of each term in the dataset. This provides an understanding of the significance of each word across all documents in the dataset. However, simply counting the term frequency does not take into account other important factors, such as the occurrence of specific terms in certain categories. As a result, it is necessary to consider additional information to accurately calculate the weight of each term.

#### G. Supervised Term Weighting Scheme

Supervised Term Weighting Schemes such as TF-MONO and its square root variant, SQRT TF-MONO was used to assign weights to each term in the dataset. This helps to improve the performance of text classification by transforming the text in documents into vectors in the vector space (Feng et al., 2018). A balance parameter  $\alpha$  with a value of 6.0 was introduced to compute the global term weight value for each term.

#### H. Machine Learning Algorithms

The researchers utilized state-of-the-art machine learning algorithms to categorize the dataset. The data was split into two (2) sets: 70% for training and 30% for testing. The researchers then applied three (3) machine learning algorithms, namely Naïve Bayes, K-Nearest Neighbors, and Linear Support Vector Classifier, in combination with supervised term weighting schemes to compare their performance. The aim was to identify which combination of algorithm and term weighting scheme yields the best results.

#### I. Evaluating the Classification Model

The classification model's performance will be evaluated by calculating precision, recall, accuracy, and F1 score.

Precision - measures how many positive predictions are correctly predicted.

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \quad (1)$$

Recall - measures how many true positive cases the classifier correctly predicted over the total number of positive case.

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \quad (2)$$

F1 Score – measures by combining both recall and precision. Also known as the harmonic mean of the two, calculating their average.

$$F1\ Score = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (3)$$

#### J. Application Simulation

The researchers chose to use Python as the primary programming language for the back-end of the study due to its interactive, object-oriented, and high-level nature. Additionally, Python offers a wide range of available modules and libraries, including the Scikit library, which makes it simple, flexible and dynamic in functionality. This made it an ideal choice for the implementation of the study, including the development of the REApp application. Furthermore, Python is a popular choice among researchers and developers because it is easy to learn, has a large community support, and is widely used in various applications such as web development, machine learning and data analysis.

#### K. ISO/IEC 25010 Software Quality

This software application will be evaluated in accordance with the ISO/IEC 25010 Software Quality standards, which provide internationally recognized guidelines for assessing the quality and performance of software systems.

## IV. RESULTS

The researchers conducted an evaluation of the effectiveness of PyTesseract OCR by selecting sample images of abstracts and backgrounds of the study from various research documents in their dataset. They performed OCR on the images and pre-processed the data to clean it.

Fig. 2 illustrates the visual representation of the abstract of a thesis titled "Safewatch: A Quarantine Symptom-Monitoring Web Application with Knowledge Discovery Using Apriori Algorithm and Naïve Bayes Classifier" and the results obtained after performing OCR on it.

Fig. 3 illustrates the visual representation of the background of the study for a capstone project titled "Barangay Information System with Decision Support System of Barangay Baan km.3 Butuan City" and the results obtained after performing OCR on it.

Fig. 4 shows the results of the pre-processing techniques employed in the extracted text of abstract and BOS. Table I illustrates the five terms with the highest frequency in the abstract dataset, along with their respective occurrences in each category.

Fig. 5 presents a word cloud comprising 200 of the most frequently occurring terms in both the Abstract and BOS datasets, arranged from left to right.

Table II presents the global factor term weights of the most frequently occurring terms in both the Abstract and BOS datasets. These values are generated from the computed results of the MONO Global method, taking into account an alpha value of 6.0. These global values will be multiplied with the term frequency if TF-MONO is to be calculated and the square root of the term frequency if SQRTF-MONO is to be calculated. It can be seen that the feature "system" has the

highest global weight of 0.029, indicating its prevalence in the abstract dataset. In contrast, the feature "custom" - which likely refers to stemmed customers - occurs more frequently in the BOS, suggesting that it is more prevalent in this dataset.

Table III displays the precision and recall scores for various combinations of term weighting schemes and machine learning algorithms used in this study for the Abstract dataset. The results indicate that the model using the squared root of TF-MONO trained with LinearSVC consistently performed the best, with a precision of 0.94, recall of 0.94, and F1 score of 0.94. However, as shown in Table 4, the SQRT TF-MONO model trained with MultinomialNB achieved slightly higher precision and recall scores, with values of 0.91 and 0.86, respectively. Additionally, the F1 score of the LinearSVC model still outperformed the classification performance of all five other trained models.

The researchers evaluated the application by conducting a survey that adhered to the ISO/IEC 25010 Standards. Twenty-five (25) participants completed the survey, comprising 15 questions divided into sections, each containing a minimum of 3 and a maximum of 4 questions. The results, displayed in Table IV, show the average percentage of responses for each

section and are rated on a scale of 5 to 1, with 5 indicating strong agreement and 1 indicating strong disagreement.

## V. DISCUSSIONS

The researchers developed an application that can classify the category of imported research papers based on four predefined categories. The application uses OCR to read text on image or pdf files and supports the classification of single or multiple research papers, and minor inaccuracies are seen in Fig. 2 and 3. Pre-processing was done in five steps to alleviate this problem, and the results are shown in Fig. 4, along with the topics extracted from the research document. The number of occurrences of each word was shown in Table I to prove the relevance of the words to each category. The results generated in Tables III and IV show that all combinations of supervised term weighting schemes and machine learning algorithms have high F1 Scores. According to the survey results, the majority voted "strongly agree" on all characteristics of ISO/IEC 25010 Standards Characteristics (see Table V). The study determined that the developed application could perform its intended functions and has met the ISO ISO/IEC 25010 Standards Evaluation Metric.

### ABSTRACT

The implementation of knowledge discovery, commonly referred to as data mining, has been considered a latent solution for the containment of the COVID-19 pandemic. The proponents developed a prototype symptom-monitoring web application as a helping tool for the concerned individuals who may have been exposed to COVID-19. By integrating Apriori Algorithm and Naive Bayes Classifier; Apriori Algorithm to mine association rules to predict the following occurring symptom based on confidence percentage, and Naive Bayes Classifier to analyze and predict the individual's risk exposure based on her/his details, the prototype system discovered that the highest risk exposure with 64% was if the individual was working on the frontline (frontliner) and the hroat are the common sequence of symptoms such as fever, cough and sore t indication with 56% confidence of risk exposure. The system aimed to address the alarming concern of being exposed to the virus due to manual monitoring, offer a platform for the individuals to track their symptoms within the entire seven-day monitoring course and produce data-driven predictions to give insights about the risk level of COVID-19 exposure. The proponents believe that with intensive research using sufficient data and the right tools for development, the prototype can become a potential tool to aid the community and healthcare organizations.

### ABSTRACT

The implementation of knowledge discovery, commonly referred to as data mining, has been considered a latent solution for the containment of the COVID-19 pandemic. The proponents developed a prototype symptom-monitoring web application as a helping tool for the concerned individuals who may have been exposed to COVID-19. By integrating Apriori Algorithm and Naive Bayes Classifier; Apriori Algorithm to mine association rules to predict the following occurring symptom based on confidence percentage, and Naive Bayes Classifier to analyze and predict the individual's risk exposure based on her/his details, the prototype system discovered that the highest risk exposure with 64% was if the individual was working on the frontline (frontliner) and the symptoms such as fever, cough and sore throat are the common sequence of symptoms with 56% confidence of risk exposure. The system aimed to address the alarming concern of being exposed to the virus due to manual monitoring, offer a platform for the individuals to track their symptoms within the entire seven-day monitoring course and produce data-driven predictions to give insights about the risk level of COVID-19 exposure. The proponents believe that with intensive research using sufficient data and the right tools for development, the prototype can become a potential tool to aid the community and healthcare organizations.

Fig. 2. Captured text image and digital text extracted thru OCR result from abstract (left-to-right image)

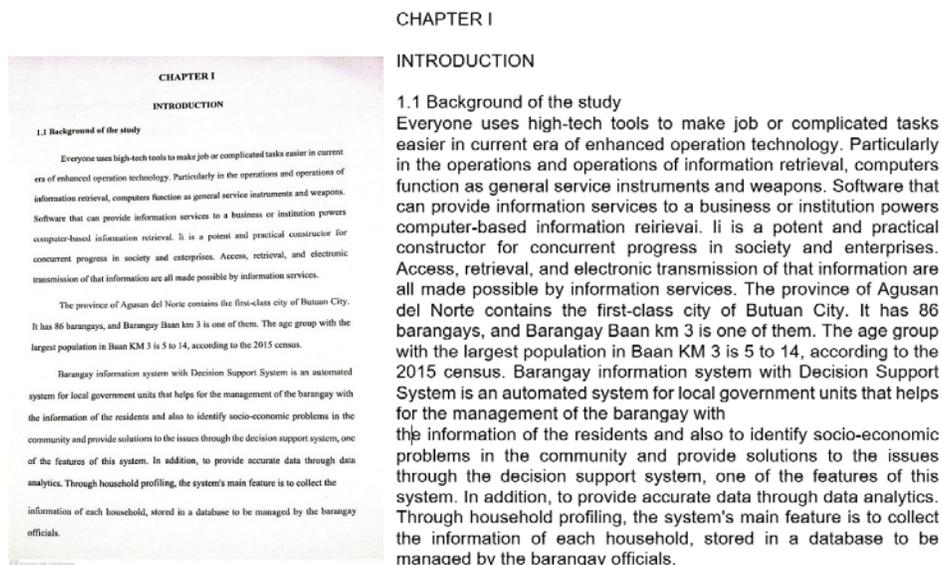


Fig. 3. Captured text image and digital text extracted thru OCR result from BOS (left-to-right image)

Stemmed

['abstract', 'abund', 'hotel', 'recommend', 'system', 'avail', 'internet', 'trivago', 'grab', 'hostel', 'zillow', 'mobil', 'problem', 'aris', 'find', 'nearbi', 'hous', 'rental', 'butuan', 'citi', 'implement', 'hous', 'rental', 'manag', 'system', 'develop', 'provid', 'optim', 'solut', 'physic', 'find', 'rental', 'hous', 'paperbas', 'process', 'manag', 'rental', 'hous', 'system', 'find', 'nearbi', 'hous', 'rental', 'avail', 'make', 'comparison', 'rental', 'hous', 'tenant', 'make', 'choic', 'accord', 'featur', 'requir', 'method', 'data', 'collect', 'collect', 'user', 'requir', 'interview', 'sourc', 'data', 'system', 'function', 'primari', 'data', 'develop', 'system', 'follow', 'waterfal', 'model', 'methodolog', 'system', 'interfac', 'written', 'design', 'use', 'python', 'clientsid', 'script', 'use', 'javascript', 'serversid', 'script', 'use', 'python', 'system', 'three', 'user', 'administ', 'owner', 'tenant', 'propon', 'care', 'analyz', 'necessari', 'data', 'gather', 'internet', 'conduct', 'survey', 'assess', 'qualiti', 'system', 'thirti', 'respond', 'compos', 'hous', 'rental', 'owner', 'twenti', 'hous', 'renter', 'assess', 'accumul', 'total', 'mean', 'equival', 'highli', 'accept', 'devic', 'inde', 'ideal', 'solut', 'help', 'landlord', 'effici', 'manag', 'inform', 'digit', 'return', 'make', 'easi', 'custom', 'search', 'housestole', 'sinc', 'help', 'landlord', 'advertis', 'rental', 'hous', 'keyword', 'hous', 'rental', 'manag', 'system', 'softwar', 'python', 'javascript', 'housestole']

Stemmed

['chapter', 'introduc', 'background', 'studi', 'everyon', 'use', 'hightech', 'tool', 'make', 'complic', 'task', 'easier', 'current', 'enhanc', 'oper', 'technolog', 'particularli', 'oper', 'oper', 'inform', 'retriev', 'comput', 'function', 'gener', 'servic', 'instrument', 'weapon', 'softwar', 'provid', 'inform', 'servic', 'busi', 'institut', 'power', 'computerbas', 'inform', 'reirievai', 'potent', 'practic', 'constructor', 'concurr', 'progress', 'societi', 'enterpris', 'access', 'retriev', 'electron', 'transmiss', 'inform', 'made', 'possibl', 'inform', 'servic', 'provinc', 'agusan', 'nort', 'contain', 'firstclass', 'citi', 'butuan', 'citi', 'barangay', 'barangay', 'baan', 'group', 'largest', 'popul', 'baan', 'accord', '2015', 'censu', 'barangay', 'inform', 'system', 'decis', 'support', 'system', 'autom', 'system', 'local', 'govern', 'unit', 'help', 'manag', 'barangay', 'inform', 'resid', 'identifi', 'socioeconom', 'problem', 'commun', 'provid', 'solut', 'issu', 'decis', 'support', 'system', 'featur', 'system', 'addit', 'provid', 'accur', 'data', 'data', 'analyt', 'household', 'profil', 'system', 'main', 'featur', 'collect', 'inform', 'household', 'store', 'databas', 'manag', 'barangay', 'offici']

Fig. 4. Pre-processed and digitized text of abstract and BOS (left-to-right image)



Fig. 5. Word cloud term frequency of abstract and BOS (left-to-right image)

TABLE I. TERM FREQUENCY DISTRIBUTION

Dataset	Term	Capstone	Thesis	Case Study	Feasibility Study	Total
Abstract	use	730	579	438	172	1919
	system	1246	415	19	30	1710
	studi	371	269	719	272	1631
	develop	680	163	135	86	1064
	research	190	188	471	140	989
Background of Study	use	657	1290	413	194	2554
	custom	128	142	2051	188	2509
	system	1278	859	128	92	2357
	product	140	137	1177	660	2114
	busi	155	100	869	828	1952

TABLE II. MONO GLOBAL WEIGHTS FOR ABSTRACT AND BOS

Term	Abstract	Term	BOS
use	1.015216258	use	1.020203602
system	1.029146199	system	1.021688587
studi	1.017633354	custom	1.032698286
develop	1.02556391	product	1.022270577
research	1.019049545	busi	1.017807377

TABLE III. EVALUATION RESULTS USING ABSTRACT DATASET

Supervised TWS with ML Models	Precision	Recall	F1 Score
TF MONO + KNeighborsClassifier	0.797330445	0.752194575	0.7299703264
TF MONO + MultinomialNB	0.576546545	0.635416668	0.703264095
TF MONO + LinearSVC	0.903361503	0.860701865	0.884272997
SQRT TF MONO + KNeighborsClassifier	0.877848725	0.844245163	0.8635014837
SQRT TF MONO + MultinomialNB	0.910247093	0.86169793	0.8872403561
SQRT TF MONO + LinearSVC	0.943446485	0.93636149	0.940652819

TABLE IV. EVALUATION RESULTS USING BOS DATASET

Supervised TWS with ML Models	Precision	Recall	F1 Score
TF MONO + KNeighborsClassifier	0.781659458	0.781659458	0.7715133531
TF MONO + MultinomialNB	0.758196003	0.758196003	0.6617210682
TF MONO + LinearSVC	0.564104215	0.564104215	0.8308605341
SQRT TF MONO + KNeighborsClassifier	0.597599638	0.597599638	0.8664688427
SQRT TF MONO + MultinomialNB	0.866617705	0.866617705	0.8902077151
SQRT TF MONO + LinearSVC	0.79807299	0.79807299	0.940652819

TABLE V. ISO/IEC 25010 STANDARDS SURVEY RESULTS FOR RECAP

ISO Characteristics	Ratings in Percentage (%)				
	5	4	3	2	1
Functional Sustainability	60%	36%	3%	1%	0%
Performance Efficiency	53.33%	40%	5.33	1.33%	0%
Usability	57%	39%	2%	2%	0%
Reliability	60%	36%	3%	1%	0%

## VI. CONCLUSION

The researchers were able to develop the classification model and application to eliminate the inconvenience and lessen the time consumption on classifying research documents for Students and Instructors. To identify which supervised term weighting scheme and machine learning algorithm can be best paired considering the evaluation metrics of each combination, the researchers used state-of-the-art supervised term weighting schemes and machine learning algorithms to simulate a classification on gathered dataset. As proven by the experiment results, the combination of SQRT TF-MONO and Linear SVC has the highest precision and recall values, and most importantly, F1 Scores of 0.94 both for the abstract and the background of the study datasets and, therefore, should be used as the classification model to classify research documents. Moreover, the researchers have developed an application prototype where users can import and classify research papers in bulk using the developed classification model. An ISO/IEC 25010 standards survey is conducted, and according to the results, most of the respondents have responded positively.

Finally, the researchers have concluded that the application can be helpful for research advisers, panelists, and reviewers

to speed up the classification time by categorizing multiple research papers at once instead of reading and manually analyzing them. Lastly, the researchers believe that the study can be improved in the future.

## ACKNOWLEDGMENT

The authors would like to express their sincere appreciation to Mr. Junell T. Bojocan of the Computer Education Department for his invaluable support in the BS Computer Science program, as well as to Mr. Gabriel Adolfo C. Malbas of the Research Extension and Innovation Department of ACLC College of Butuan for his generous financial support. The authors recognize the instrumental role of their contributions in completing this study.

## REFERENCES

- [1] X. Luo, "Efficient English text classification using selected Machine Learning Techniques," *Alexandria Eng. J.*, vol. 60, no. 3, pp. 3401–3409, 2021, doi: 10.1016/j.aej.2021.02.009.
- [2] S.-W. Kim and J.-M. Gil, "Research paper classification systems based on TF-IDF and LDA schemes," *Human-centric Comput. Inf. Sci.*, vol. 9, no. 1, p. 30, 2019, doi: 10.1186/s13673-019-0192-7.

- [3] D. Sarkar, *Text Analytics with Python - A Practitioner's Guide to Natural Language Processing*. 2019.
- [4] A. Allahverdi-pour and F. S. Gharehchopogh, "An Improved K-Nearest Neighbor with Crow Search Algorithm for Feature Selection in Text Documents Classification," *J. Adv. Comput. Res.*, vol. 9, pp. 37–48, 2018.
- [5] G. B. Holanda et al., "Development of OCR system on android platforms to aid reading with a refreshable braille display in real time," *Meas. J. Int. Meas. Confed.*, vol. 120, pp. 150–168, 2018, doi: 10.1016/j.measurement.2018.02.021.
- [6] H. Goodrum, K. Roberts, and E. V. Bernstam, "Automatic classification of scanned electronic health record documents," *Int. J. Med. Inform.*, vol. 144, no. June, p. 104302, 2020, doi: 10.1016/j.ijmedinf.2020.104302.
- [7] C. C. Abalorio and M. Cerna, "Course Evaluation Generator (Ceg): An Automated Academic Advising System with Optical Character Recognition," *Int. J. Technol. Eng. Stud.*, vol. 4, no. 5, pp. 189–196, 2018, doi: 10.20469/ijtes.4.10003-5.
- [8] I. Alsmadi, "Term weighting scheme for short-text classification: Twitter corpuses," *Neural Comput. Appl.*, vol. 8, 2018, doi: 10.1007/s00521-017-3298-8.
- [9] M. Raza, F. K. Hussain, O. K. Hussain, M. Zhao, and Z. ur Rehman, "A comparative analysis of machine learning models for quality pillar assessment of SaaS services by multi-class text classification of users' reviews," *Futur. Gener. Comput. Syst.*, vol. 101, pp. 341–371, 2019, doi: 10.1016/j.future.2019.06.022.
- [10] X. Zhou et al., "A survey on text classification and its applications," *Web Intell.*, vol. 18, no. 3, pp. 205–216, 2020, doi: 10.3233/WEB-200442.
- [11] Q. Li et al., "A Survey on Text Classification: From Shallow to Deep Learning," *arXiv*, 2020, doi: 10.48550/ARXIV.2008.00364.
- [12] Muna Ahmed Awel and A. I. Abidi, "Review on Optical Character Recognition," *Int. Res. J. Eng. Technol.*, vol. 6, no. 6, pp. 3666–3669, 2019, [Online]. Available: [www.irjet.net](http://www.irjet.net)
- [13] V. Z. V Singco, J. C. Trillo, C. C. Abalorio, J. C. M. Bustillo, J. T. Bojocan, and M. C. Elape, "OCR-based Hybrid Image Text Summarizer using Luhn Algorithm with FinetuneTransformer Models for Long Document," vol. 13, no. 02, 2023, doi: 10.46338/ijetae0223.
- [14] P. Divya et al., "Web based optical character recognition application using flask and tesseract," *Mater. Today Proc.*, 2021, doi: 10.1016/j.matpr.2020.10.850.
- [15] T. Dogan and A. K. Uysal, "On Term Frequency Factor in Supervised Term Weighting Schemes for Text Classification," *Arab. J. Sci. Eng.*, 2019, doi: 10.1007/s13369-019-03920-9.
- [16] Z. Tang, W. Li, and Y. Li, "An improved supervised term weighting scheme for text representation and classification," *Expert Syst. Appl.*, vol. 189, p. 115985, 2022, doi: <https://doi.org/10.1016/j.eswa.2021.115985>.
- [17] T. Dogan and A. K. Uysal, "A novel term weighting scheme for text classification: TF-MONO," *J. Informetr.*, vol. 14, no. 4, p. 101076, 2020, doi: 10.1016/j.knosys.2012.06.005.
- [18] C. C. Abalorio, R. P. Medina, A. M. Sison, and G. A. Dalaorao, "Extended Max-Occurrence with Normalized Non-Occurrence as MONO Term Weighting Modification to Improve Text Classification," *Int. J. Adv. Comput. Sci. Appl.*, vol. 13, no. 4, pp. 91–97, 2022, doi: 10.14569/IJACSA.2022.0130411.
- [19] C. C. Abalorio, A. M. Sison, R. P. Medina, and G. A. Dalaorao, "Applying EMONO Variants to Multi-Class Sentiment Analysis for Short-Distance Inter-Class Frequency of Term," vol. 71, no. 4, pp. 1938–1947, 2022.
- [20] A. Basarkar, "Document Classification using Machine Learning," 2017, doi: <https://doi.org/10.31979/etd.6jmu-9xdt>.
- [21] A. Barua, O. Sharif, and M. M. Hoque, "Multi-class Sports News Categorization using Machine Learning Techniques: Resource Creation and Evaluation," *Procedia Comput. Sci.*, vol. 193, pp. 112–121, 2021, doi: <https://doi.org/10.1016/j.procs.2021.11.002>.
- [22] Y. HaCohen-Kerner, D. Miller, and Y. Yigal, "The influence of preprocessing on text classification using a bag-of-words representation," *PLoS One*, vol. 15, no. 5, pp. 1–22, 2020, doi: 10.1371/journal.pone.0232525.
- [23] A. I. Kadhim, "Survey on supervised machine learning techniques for automatic text classification," *Artif. Intell. Rev.*, vol. 52, no. 1, pp. 273–292, 2019, doi: 10.1007/s10462-018-09677-1.
- [24] J. C. M. Bustillo, R. P. Medina, A. M. Sison and M. Y. Orong, "Predictive Hybridization Model integrating Modified Genetic Algorithm (MGA) and C4.5," 2022 6th International Conference on Electronics, Communication and Aerospace Technology, Coimbatore, India, 2022, pp. 1500-1507, doi:10.1109/ICECA55336.2022.10009532.

# R-Diffset vs. IR-Diffset: Comparison Analysis in Dense and Sparse Data

Julaily Aida Jusoh, Sharifah Zulaikha Tengku Hassan, Wan Aezwani Wan Abu Bakar,  
Syarilla Iryani Ahmad Saany, Mohd Khalid Awang, Norlina Udin @ Kamaruddin  
Faculty of Informatics and Computing, University Sultan Zainal Abidin, Besut, Terengganu, 22200, Malaysia.

**Abstract**—The mining of concealed information from databases using Association Rule Mining seems to be promising. The successful extraction of this information will give a hand to many areas by aiding them in the process of finding solutions, economic projecting, commercialization policies, medical inspections, and numbers of other problems. ARM is the most outstanding method in the mining of remarkable related configurations from any groups of information. The important patterns encountered are categorized as recurrent/frequent and non-recurrent/infrequent. Most of the previous data mining methods concentrated on horizontal data set-ups. Nevertheless, recent studies have shown that vertical data formats are becoming the main concerns. One example of vertical data format is Rare Equivalence Class Transformation (R-Eclat). Due to its efficacy, R-Eclat algorithms have been commonly applied for the processing of large datasets. The R-Eclat algorithm is actually comprised of four types of variants. However, our work will only focus on the R-Diffset variant and Incremental R-Diffset (IR-Diffset). The performance analysis of the R-Diffset and IR-Diffset algorithms in the mining of sparse and dense data are compared. The processing time for R-Diffset algorithm, especially for sequential processing is very long. Thus, the incremental R-Diffset (IR-Diffset) has been established to solve this problem. While R-Diffset may only process the non-recurrent itemsets mining process in sequential form, IR-Diffset on the other hand has the capability to execute sequential data that have been fractionated. The advantages of this newly developed IR-Diffset may become a potential candidate in providing a time-efficient data mining process, especially those involving the large sets of data.

**Keywords**—R-Diffset; IR-Diffset; dense data; sparse data; comparison analysis

## I. INTRODUCTION

The main objective of data mining (DM) is the process of identifying patterns for useful information in large data repositories [1-3, 36, 26, 30]. It plays a key role in the oversized data approach to acquire meaningful knowledge from a complex system. The approach includes collaborative studies in statistics, machine learning, data science, and database theory. Its goal is to learn about the past or predict the future by studying the present data. In addition, this approach, also known as "Knowledge Discovery in Databases" (KDD), focuses on finding patterns in databases, resulting in an association rule that may disclose important information. Some common patterns are found in the databases, such as clusters, sets of items, trends, and outliers [4]. In data mining, there are two types of tasks: prediction tasks and descriptive tasks. The prediction tasks attempt to determine the value of one attribute

depending on the value of another. These tasks incorporate techniques like statistics, categorization, regression, and forecasting [19]. Meanwhile, the descriptive tasks are to generate patterns in the database in order to extract the underlying relationships [29]. Some of the patterns generated are anomalies, clusters, correlations, and trends. Clustering, summarizing, association rules, pattern identification, and sequence discovery are some of the approaches used in this task [22]. There are two types of itemset mining in the database: frequent itemset and infrequent itemset. From the records of the previous studies, in frequent itemset mining, there are three well-known algorithms: Apriori [5, 6, 32], FP-Growth [7], and ECLAT [8, 18]. This research focuses on the vertical format by looking deeper into the equivalence class transformation (ECLAT) algorithm [8]. Tidset, Diffset, Sortdiffset, and Postdiffset are four extension variations introduced in ECLAT. In 2018, Jusoh et al. [9-10] introduced the R-ECLAT algorithm, particularly useful for mining infrequent itemsets. This algorithm is based on the ECLAT algorithm for mining infrequent itemsets [11-13]. In previous research, the R-Diffset was developed and executed in sequential processing for infrequent item mining as an extension of the Diffset algorithm. However, the current execution of data processing often faces the constraint of slowness, especially when dealing with large datasets.

In 2022, Man et al. [19-20, 27, 43] introduced a new incremental of rare pattern mining approach using R-Eclat named as Incremental Rare Equivalence Class Transformation, (IR-Eclat). The IR-Eclat, which was created especially for mining rare patterns is advantageous for dynamic databases because the volume of data increase linearly with time. For dynamic databases that are subject to the addition or deletion of items or records of transactions in the database, the incremental approach is advantageous. To apply mined knowledge of previously and scanning of a basic incremented database, it is necessary to use implementation of incremental mining.

An incremental approach is a new approach that has been executed sequentially on the splitted data. This incremental approach will complement the IR-Diffset algorithm to ensure that it can mine the infrequent itemsets faster. This new approach is introduced as IR-Diffset, where IR represents incremental-rare. As part of the R-Diffset algorithm, this study introduces an incremental approach to deal with the processing time issue. To simplify, this research is significant in determining the following aspects: How to reduce speedy processing time in mining infrequent pattern especially in huge dataset? The rest of the components are then organized as

follows: Section IIA discusses sparse and dense data. The basics of mining rare items are covered in Section IIB. R-Diffset and IR-Diffset are described in Section III. Section IV clarifies the discussion of the study. The study's results and the discussion are also described in Section IV. The research was concluded in Section V.

## II. RELATED WORKS

This section concerned on discussion regarding database, data mining and infrequent itemset mining. A database is known as a structured collection of data, dataset or record. It can be effortlessly rapid accessed, managed and updated data in conjunction with various data processing operations. Data mining is the process of extracting useful unknown knowledge from large datasets. However, mining infrequent (rare) itemsets may be more interesting in many real-life applications. However, this section also covers about sparse and dense data.

### A. Sparse vs Dense Data

Sparse data is a matrix [14] in numerical analysis, where the majority of the elements are zero. This section will discuss about sparse data, which are quicker to be processed and figured, compared to dense data. Although the number of zero-valued elements are not specified for a sparse matrix, the number of non-zero elements are standardized to be equal to the number of rows or columns, due to their similarities. However, if most of the elements are non-zero, the matrix will be identified as dense. The sparsity of a matrix is calculated by dividing the zero value elements with those of non-zeros (e.g.,  $p \times q$  for a  $p \times q$  matrix) [15]. The sparse matrix can be described in the following ways: array and linked list representation. In the array representation, the 2D array can be used to represent a sparse matrix in which there are three elements named as follows:

- Row: a row index containing a non-zero element.
- Column: a column index containing a non-zero element.
- Value: The non-zero element's value is placed at the index (row, column).

A sparse matrix is applied in linked list to portray the list of data structures. While there are three elements in the array representation (i.e., row, column, and value), the linked list representation on the other hand, comprises of four elements. The elements in the linked list are as follow:

- Row: a row index containing a non-zero element.
- Column: a column index containing a non-zero element.
- Value: The non-zero element's value is placed at the index (row, column).
- Next node: It records the next node's address.

When storing and processing sparse matrices on a computer, it is preferable to utilize specific algorithms and data structures that take advantage of the matrices' sparse structure. Sparse matrices, which are common in machine learning, have given rise to specialized computers. When applied to vast sparse matrices, typical dense-matrix structures and techniques

are sluggish and ineffective, as computation and storage are squandered on the zeros.

Sparse data compresses more effectively and hence take up less storage space. Some outsize sparse matrices are unmanageable using typical dense-matrix techniques.

The matrix *sparsity* is like density of zero elements in the sparse matrix and is defined as the number of zero elements divided by total elements in the matrix, which computed as:

$$\text{sparsity} = \frac{\text{number of zero elements}}{\text{number of total elements}}$$

The sparsity of a sparse matrix is always greater than 0.5.

In a dense matrix, most of the elements are non-zero. The dense matrix can be stored in a fixed-size array.

The sparse matrix, on the other hand, recorded only non-zero elements in a dictionary mapping an index  $(i, j)$  to its entry. As a result, you should only describe a matrix as sparse if the number of non-zero elements is relatively tiny in comparison to the total number of entries. This is generally advantageous if the non-zero elements are large (there are about  $n$  non-zero entries compared to about  $n^2$  zero entries). However, generating dense matrices is usually faster for extremely tiny matrices (e.g.,  $5 \times 5$ ), even with a large number of zero entries.

Fig. 1 depicts the characteristics of sparse and dense data. Dense data is the nature of the data itself, which is dense, which causes it to be slow. The data is dense and takes up a lot of space, and it is extremely difficult to process. In order to find other data using the dense data, it must first check each column one by one; without skipping a column. Sparse data indicates the nature of the data itself, which is sparse. Processing time for sparse data is faster compared to dense data because sparse data is not dense or compact; in the sparse data column, there is still an empty space, so processing the data will be faster and will not glitch in comparison to the dense data.

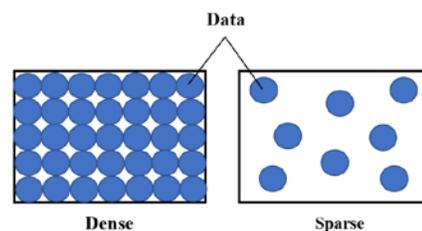


Fig. 1. Features of data-sparse and data-dense.

The theoretical background for sparse and dense data has been thoroughly discussed in this section. The next section will focus on one of the infrequent itemset mining.

### B. Infrequent Item Set Mining

Up to this date, the mining of non-recurrent data have been considerably developed. Non-recurrent or infrequent patterns are usually utilized in many disciplines such as biology, health, and security. In medical sector, the analysis of clinical conditions of patients are carried out to detect irregular patterns

or trends. The results will help the medical officers in deciding the needed treatments or prescriptions.

Infrequent itemset mining from a database is not as prominent as frequent itemset mining [39-40]. However, there are few areas where infrequent item mining is more applicable than frequent item mining [16]. The objective of IIM is to discover unusual, but informative, relationships between entries in a dataset. In [16] the concept of itemset for frequent pattern mining, were established based the pattern mining concept. The set of items were denoted as  $I = \{I_1, I_2, \dots, I_m\}$ . The itemset  $I$  will be accounted as frequent if and only if the occurrence frequency in the database equals or exceeds the minimum support threshold defined by the users [24-25]. In contrast, infrequent itemset mining requires that the frequency of occurrence in the database be equivalent to or lesser than the user-defined minimum support threshold. But, if the threshold is too low, infrequent itemset mining using the traditional frequent itemset mining method could be unsuccessful [33].

Infrequent itemset mining are receiving tremendous concern principally in the fields of networking and medicine. In networks, it is used to investigate any atypical incidences in any networks which are usually indicators of network breakdown or security breach. In medical, infrequent itemset mining may assist for the discovery of treatments for uncommon cases. The objective of infrequent itemset mining is to mine patterns with a support value smaller than the minimum threshold [34, 35]. Yet, the process of extracting rare patterns from the database, appears to be challenging.

In 2019, Man et al. [4, 28] proposed algorithm to address the discovery of infrequent itemsets mining from the transactional database based on Eclat algorithm. In support measure, IF-Diffset was proved to perform better upon encountering the infrequent itemsets of the transactional database.

An algorithm for extracting infrequent itemsets from weblog has been introduced by Bakariya et al. in 2019 [17]. Infrequent Itemset Mining for Weblog (IIMW) algorithm is a phased, top-down, broad-first algorithm for the recovery of infrequent item sets. A power set and lattice traversal approach had been utilized for this purpose. This approach followed the top-down mechanism, begun from top and carried out down to the bottom. By applying this approach, the computation of support for smaller itemsets were simpler than the larger itemset. This problem has been acknowledged as the challenges for the further application of this method.

To overcome the above-mentioned obstacles, Lu et al., later in 2020, [42] suggested an infrequent pattern mining algorithm using a top-down and depth-first traversing strategy. A negative itemset tree was applied to speed up the mining procedure, by compressing the datasets for a quicker counting process. Their itemset miner, denoted as NIIMiner had been verified to solve the problems of rare itemset mining.

In 2021, Darrab et al. [21] performed an extensive search of latest methods of rare itemset mining. It has been done by inspecting another efficient data structure with mining rare itemsets, the mining of most interesting rare itemsets, as well

as the mining rare of itemsets with multiple minimum support thresholds.

In 2021, Abu bakar and his fellow researchers [31, 37] developed a performance enhancement in Incremental Eclat (iEclat) model by embedding Critical Relative Support (CRS) in the mining of infrequent itemset. This was done in line to the increment of itemsets which produced higher cardinality of intersection between each item hence required the method of vertical mining.

In 2022, Cui et al. [41] introduced a novel fuzzy-based rare itemset mining algorithm called FRI-Miner, and successfully discovered valuable and interesting fuzzy rare itemsets in a quantitative database by applying fuzzy theory with linguistic meaning. The established algorithm was a success and has been shown to have an improved overall mining quality compared to the existing algorithm.

The previous paragraph of this section summarized the latest algorithm for infrequent itemset mining which are infrequent itemsets mining from the transactional database based on Eclat algorithm [4], Infrequent Itemset Mining for Weblog (IIMW) algorithm [17], Incremental Eclat (iEclat) model by embedding Critical Relative Support (CRS) in mining of infrequent itemset [31], Infrequent Pattern Mining Using Negative Itemset Tree (NIIMiner) algorithm [42] and Fuzzy-based Rare Itemset Mining (FRI-Miner) algorithm [41]. The next section will pay a focus on R-Diffset and IR-Diffset. The constraint of time, due to the large size of data, is among the main disadvantages often faced by the R-Diffset. For improvements, a new algorithm, known as incremental R-Diffset (IR-Diffset), which could help to fasten the process has been introduced.

### III. R-DIFFSET VS. IR-DIFFSET

This section will explain the theoretical background of R-Diffset and IR-Diffset. The technique, formula, and illustration model for R-Diffset and IR-Diffset in infrequent itemsets mining will also be presented. Similar techniques, called depth-first search are used for both R-Diffset and IR-Diffset.

#### A. R-Diffset

In [4], the extension of the diffset algorithm named as R-Diffset, has been introduced. It is executed via sequential processing specifically for infrequent itemset mining. Each item in a vertical database is associated with its corresponding tidset, i.e., the set of all transactions (tids) where it occurs. Thus, the size of tidsets is the primary factor affecting the running time and memory usage. The bigger the tidsets, the more time and memory are needed. However, existing data processing execution is often confronts the issue of large database with single format only. Before this, the execution time data processing in R-Diffset algorithm it is very time-consuming especially in sequential processing [38].

The R-Diffset algorithm concern to keeps track the differences in the tids (transaction IDs) of a candidate pattern from its generating infrequent patterns. It helps to decrease the size of memory required to store intermediate results as depicted in the below pseudocode. The line 6 till 8 in Fig. 2 illustrate that this algorithm will generate all itemsets with their

diffsets and supports. Line number 7 shows where it gets the diffset data, instead of getting intersection of tidsets data between column  $i$  and  $(i + 1)$  based on minimum support that has been set at line 5. All infrequent itemsets are generated by computing diffsets for all distinct pairs of itemsets and checking the support of the resulting itemset. This process is repetitive until all infrequent itemsets have been enumerated. Since the generated diffsets are a small fraction size of tidsets, so the intersection operations are performed extremely fast. The performance of R-Diffset shows a good improvement in execution time over R-Tidset variant in both dense and sparse databases.

Pseudocode R-Diffset
Input: $E((i_1, t_1), \dots, (i_n, t_n) P), s_{min}$ Output: $IF(E, s_{min})$
Begin //get minimum support 1. arrange data by itemset 2. looping = numOfColumns; 3. min_supp = number_of_rows * percentage_min_support 4. run tidset; 5. for (i=0; i<=min_support) 6. if (support<=min_support){ 7. get diffset data for column [i] with column [i+1]; 8. save to DB;} 9. Set next transaction data; 10. Write to text file the value for the current / last transaction data;} end

Fig. 2. Pseudocode for R-Diffset.

### B. IR-Diffset

This study presents a modified algorithm, which is an improvised version of existing R-Diffset, named as Incremental R-Diffset (IR-Diffset). This new approach serves as an alternative to speed up the processing time in infrequent itemsets mining, particularly those involving large datasets. IR-Diffset could be implemented with the data that has been sequentially splitted, incrementally. Incremental in itemsets refers to any additional new item which is being added or deleted to the existing itemsets in database whereas incremental in records of transaction refers to the additional transactions, added to the existing transaction. The IR-Diffset model is illustrated in Fig. 3.

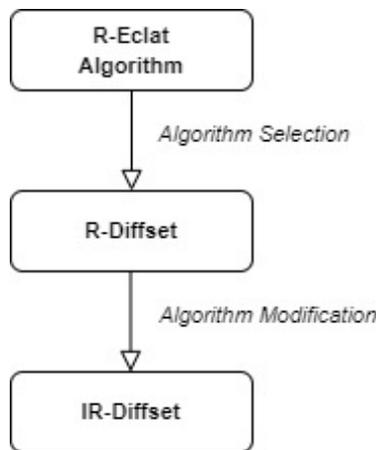


Fig. 3. IR-Diffset Model.

The incremental itemset will be added to the basic model of R-Diffset, establishing a new IR-Diffset model. The IR-Diffset algorithm will keep track the differences in the tids (transaction IDs) by counting each item for finding support which is less than or equal. As presented in Fig. 4, a slight modification of implemented code is at line 7 (while loop). The while loop functions to control flow statements by allowing itemsets to be executed repeatedly. For example, the itemsets in the loop will be performed repeatedly, as long as the variable  $i$  is less than looping. The variable ' $r$ ' at line 4, refers to the record of transaction. When variable ' $r$ ' is not equal to zero, the transactions will be run (i.e the condition is true) it will enter the looping. When variable ' $r$ ' is equals to zero, the process will be terminated. Then, line 8 refers to  $(i = 0; i < looping; i++)$ , which is  $i = 0$  means that the variable  $i$  is = to 0,  $i < looping$  means that the process will be run repeatedly as long as  $i$  is smaller than zero and  $i++$  means that each time the looping process take place, the new  $i$  will be  $i = i + 1$ . This incremental approach has been proposed to overcome the limitations of sequential processing in extracting infrequent itemsets. Then, the following are the key steps in IR-Diffset over the dataset:

Support counting is for finding the support through determination of supports of any  $k$ -itemsets on the intersecting tid-lists of its  $k-1$  subsets. First, the basic theory for support counting must be less than or equal to MSTV. Second, the result will be written and saved in the text file prior to preparation for the next row of transaction data. The minimum support threshold value (MSTV) is considered as a benchmark to discover a low occurrence in each dataset. MSTV is determined in terms of percentage [23].

$$\frac{\alpha}{100} * \beta \tag{1}$$

Where,  $\alpha$  = User specified minimum support value  
 $\beta$  = Total of records in datasets.

Pseudocode IR-Diffset
Input: $E((i_1, t_1), \dots, (i_n, t_n) P), s_{min}$ Output: $IF(E, s_{min})$
start //get minimum support sort data by itemset looping = num_of_column r = record_of_transactions min_supp = num_of_row * percentage_min_supp run tidset while r ≠ 0 do for (i=0; i<looping; i++) if(support<=min_support) get diffset data for column [i] with column [i+1]; add next transaction data; write to text file the value for the current / last transaction data; end

Fig. 4. Pseudocode for IR-Diffset.

The step-by-step process of IR-Diffset is presented in the above pseudocode. Other process steps in each of these variants are preserved as the original variant. The first dissimilarity is the basic theory for support counting where it will only consider the support that is less than or equal to MSTV. The second difference is the storage of results. The

value/result is written and saved in the text file prior to preparation for the next row of transaction data. This is for the purpose of saving memory space. In each loop, starting with the first loop, if the support is less than or equal ( $\leq$ ) to  $min\_supp$ , then,

1) Rather than utilizing intersection, IR-Diffset obtains the result of Diffset (difference intersection set) between the  $k^2$  and  $k^{th}+1$  columns and stores it in the database.

Table I shows the effectiveness of the comparative analysis and that clearly shows the technique, data format, execution time, process, and execution between R-Diffset and IR-Diffset algorithms with infrequent itemsets mining. R-Diffset and IR-Diffset use the same technique, depth-first search.

	R-Diffset	IR-Diffset
<b>Technique</b>	Depth first search	
<b>Data format</b>	Vertical	
<b>Time</b>	Faster intersection process	Less fast for intersection process
<b>Process</b>	Keeps track of differences in tidsets	-
<b>Execution</b>	Executes the infrequent itemsets mining process in sequential form	Executes sequentially on the data that has been split

Table I summarizes the differences of R-Diffset and IR-Diffset. The details of step-by-step processes for both R-Diffset and IR-Diffset are previously presented in Fig. 2 and Fig. 4, respectively. Line 6 to 8 in Fig. 2 illustrate that all itemsets will be generated by this algorithm together with their diffsets and supports. This algorithm is similar to those of IR-Diffset as portrayed in Fig. 4 (lines 9 to 11). The differences between these two pseudocodes can be seen in line 7 (while  $r \neq 0$ ) is used inside the pseudocode of the IR-Diffset. The term "min\_supp" is used in the pseudocode to refer to the minimum support threshold value, which is expressed as a percentage and is calculated by dividing the user-specified min\_supp value by 100 and multiplying it by the total number of rows (records) in each dataset. Then, starting with the first loop, if the support is less than or equal ( $\leq$ ) to min\_supp, then line number 10 shows where it gets the diffset data, instead of getting intersection of tidsets data between column  $i$  and  $(i + 1)$  based on minimum support that has been set at line 8 and saved to DB. The benefit of incremental storage structure lies in the fact that every candidate of itemsets in search space has the same itemset in database. Due to this, its support can be computed by a few simple database operations, removing the need of full scan of database. The rest of pseudocode lines for R-Diffset and IR-Diffset are same.

#### IV. DISCUSSION

All experimentations are carried out on an HP Notepad with an Intel® Core™ i7-3520M CPU running at 2.90 GHz and 8GB of RAM running Windows 10 64-bit. Open-

source software is used to implement the algorithm development software standard. For our database server, we use MySQL (version 5.6.25 - MySQL community server (GPL)), Apache/2.4.16 (Win32) OpenSSL/1.0.1i PHP/5.6.11, PHP as a programming language, and phpMyAdmin (version 4.8.4) to manage MySQL via the Web are all included in the package. Four datasets including chess, pumsb\_star, retails, and T40110D100K are used in this experimentation as described in Table II. The benchmark datasets were collected in their raw form from the Frequent Itemset Mining Dataset Repository (<http://fimi.ua.ac.be/data/>). The benchmark is also converted into Structured Query Language (SQL) format to make things easier.

Datasets	No. of Transactions	Length (Attributes)	Size (KB)	Category
Chess	3196	37	335	Dense
Pumb_star	49046	57	11526	Dense
Retails	88162	68	5143	Sparse
T40110D100K	100001	32	15116	Sparse

In order to facilitate and accelerate research, the datasets are limited to a thousand rows of randomly processed item sets for mining purposes. Our research is focused on R-Diffset and IR-Diffset. Fig. 5 depicts the performance assessment graph in terms of execution time (in seconds) for the four (4) datasets: chess, pumsb\_star, retails and T40110D100K.

The experiments on R-Diffset and IR-Diffset is successfully tested to determine the comparative performance between sparse and dense data. Fig. 5 summarizes that the execution time R-Diffset in dense data is 0.0001 slightly faster than IR-Diffset. But, in sparse data, the result shows that the execution time of IR-Diffset is 0.0003 faster over R-Diffset.

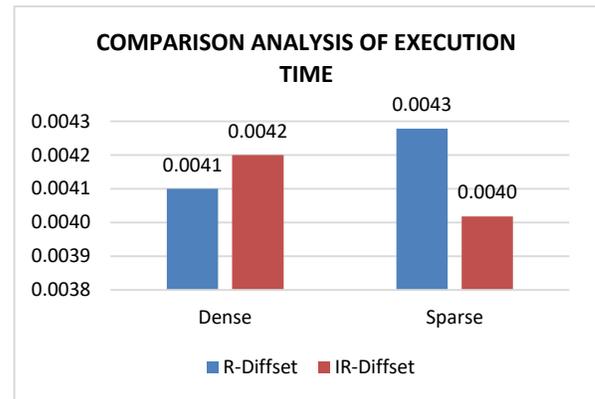


Fig. 5. Performance on R-Diffset and IR-Diffset for chess, pumsb\_star, retails, and T40110D100K

The difference of execution time for both algorithms may be caused by the property's nature of the data itself. Dense dataset is naturally looks like a compressed data where the space between the data is very small. Due to this characteristic, the process of mining infrequent itemset via incremental approach become slower when it needs to sequentially mining the infrequent itemset in several fractions. Finding every single of itemset in the dense data is incredibly challenging due to its dense characteristic. Nevertheless, it is very contrary compared

to the nature feature in sparse dataset. In sparse dataset, the feature between the data itself is sporadic and the space between each data is wide. Due to the wide space between each data, the discovery of the infrequent itemsets is a bit faster and less challenging.

## V. CONCLUSION

In this research, the performance comparison between R-Diffset and IR-Diffset has been analysed. The IR-Diffset have a better performance for infrequent itemset mining in sparse dataset although it slightly loses its advantages over R-Diffset in dense dataset. The incremental approach is introduced to overcome the limitations of sequential processing in extracting infrequent itemsets. The complementary of this approach is believed can be a solution of speedy processing time in mining infrequent pattern especially in huge dataset. Looking at the improvement performance of IR-Diffset in sparse dataset, it can be applied to the real dataset from various fields, such as finding a rare disease at certain areas. It might help on how to control and manage this disease from dispersion which might causes the increase number of patients.

## ACKNOWLEDGMENT

We would like to express our gratitude to the CREIM of UniSZA for providing financial support under UniSZA internal grant code (UniSZA/2021/GOT/02). Thanks to the corresponding authors, Dr. Julaily Aida Jusoh as the CREIM-UniSZA research project leader, Sharifah Zulaikha Tengku Hassan as the research assistant, and the grant participants. We would also like to thank all of the faculty members who helped us evaluate spelling problems and synchronization consistency and for their thoughtful remarks and recommendations.

## REFERENCES

- [1] Fayyad, U. M., Piatetsky-Shapiro, G., & Smyth, P. (1996, August). Knowledge Discovery and Data Mining: Towards a Unifying Framework. In *KDD* (Vol. 96, pp. 82-88).
- [2] Gullo, F. (2015). From patterns in data to knowledge discovery: What data mining can do. *Physics Procedia*, 62, 18-22.
- [3] Tan, P. N., Steinbach, M., & Kumar, V. (2006). Introduction to data mining, Addison Wesley publishers.
- [4] M. Man, J. A. Jusoh, S. I. A. Saany, W. A. W. A. Bakar, and M. H. Ibrahim, "Analysis Study on R-ECLAT Algorithm in Infrequent Itemsets Mining", *International Journal of Electrical and Computer Engineering (IJECE)*, pp. 5446-5453, 2019.
- [5] RAgrawal R, Imieliński T, Swami A., "Mining association rules between sets of items in large databases". *Proceedings Acm sigmod record*, vol. 22, no. 2, pp. 207-216, 1993.
- [6] Agrawal, R., & Srikant, R. (1994, September). Fast algorithms for mining association rules. In *Proc. 20th int. conf. very large data bases, VLDB* (Vol. 1215, pp. 487-499).
- [7] Han J, Pei J, Yin Y. "Mining frequent patterns without candidate generation," *Proceedings ACM sigmod record*, vol. 29, No. 2, pp. 1-12, 2000.
- [8] Ogihara Z. P., Zaki M. J., Parthasarathy S., Ogihara M., Li W., "New algorithms for fast discovery of association rules," 3 rd Intl. Conf. on Knowledge Discovery and Data Mining, 1997.
- [9] Jusoh, J. A., Man, M., & Bakar, W. (2018). Mining infrequent patterns using R-ECLAT algorithms. *J Fundam Appl Sci*, 24.
- [10] Jusoh, J. A., & Man, M. (2018). Modifying iEclat Algorithm for Infrequent Patterns Mining. *Advanced Science Letters*, 24(3), 1876-1880.
- [11] Zaki, M. J., Parthasarathy, S., Ogihara, M., & Li, W. (1997, August). New algorithms for fast discovery of association rules. In *KDD* (Vol. 97, pp. 283-286).
- [12] Zaki, M. J., & Gouda, K. (2003, August). Fast vertical mining using diffsets. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 326-335).
- [13] Trieu, T. A., & Kunieda, Y. (2012, February). An improvement for dECLAT algorithm. In *Proceedings of the 6th International Conference on Ubiquitous Information Management and Communication* (pp. 1-6).
- [14] El Abbadi, N. K., & Al Tae, E. J. (2014). An Efficient Storage Format for Large Sparse Matrices based on Quadtree. *International Journal of Computer Applications*, 105(13).
- [15] Bik, A. J., & Wijshoff, H. A. (1993, August). Compilation techniques for sparse matrix computations. In *Proceedings of the 7th international conference on Supercomputing* (pp. 416-424).
- [16] Borah, A., & Nath. "Rare Pattern Mining: Challenge and Future Perspectives." Springer International Publishing. (2018).
- [17] Bakariya, B., Thakur, G. S., & Chaturvedi, K. (2019). An efficient algorithm for extracting infrequent itemsets from weblog. *Int. Arab J. Inf. Technol.*, 16(2), 275-280.
- [18] M. J. Zaki and K. Gouda, "Fast vertical mining using diffsets," *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, pp. 326-335, 2003, doi: 10.1145/956750.956788.
- [19] Man, M., Ruslan, N. A. B., Bakar, W., Jusoh, J. A., Yusof, M. K., & Josdi, N. L. N. B. (2022). IR-ECLAT: A New Algorithm for Infrequent Itemset Mining. *Journal of Theoretical and Applied Information Technology*, 100(11).
- [20] Man, M., Ruslan, N. A., Jusoh, J. A., & Bakar, W. A. W. A. (2020). Conceptual Model of Incremental R-Eclat Algorithm for Infrequent Itemset Mining. *International Journal Of Engineering Trends And Technology (IJETT)*, 10, 129-133.
- [21] Darrab, S., Broneske, D., & Saake, G. (2021). Modern applications and challenges for rare itemset mining. *Int J Mach Learn Comput*, 11(3), 208-218.
- [22] Fournier-Viger, P., Lin, J. C. W., Vo, B., Chi, T. T., Zhang, J., & Le, H. B. (2017). A survey of itemset mining. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 7(4), e1207.
- [23] W. A. B. W. A. Bakar, et al., "Incremental-Eclat Model: An Implementation via Benchmark Case Study," Springer International Publishing Switzerland, P.J. Soh et al. (eds.), *Advances in Machine Learning and Signal Processing, Lecture Notes in Electrical Engineering*, vol. 387, pp. 35-46, 2016.
- [24] Vu, L., & Alaghand, G. (2014, January). An efficient approach for mining association rules from sparse and dense databases. In *2014 World Congress on Computer Applications and Information Systems (WCCAIS)* (pp. 1-8). IEEE.
- [25] Darrab, S., & Ergenc, B. (2017). Vertical Pattern Mining Algorithm for Multiple Support Thresholds. *Procedia Computer Science*. 112. 417-426. Elsevier.
- [26] M. K. Yusof and M. Man, "Efficiency of JSON for Data Retrieval in Big Data," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 7, no. 1, pp. 250-262, Jul. 2017, doi: 10.11591/IJEECS.V7.I1.PP250-262.
- [27] Man, M., Bakar, W. A. W. A., Jalil, M. M. A., & Jusoh, J. A. (2018). Postdiffset Algorithm in Rare Pattern: An Implementation via Benchmark Case Study. *International Journal of Electrical and Computer Engineering (IJECE)*, 8, 4477-4485.
- [28] Bakar, W. A. W. A., Jalil, M. A., Man, M., Abdullah, Z., & Mohd, F. (2018). Postdiffset: an Eclat-like algorithm for frequent itemset mining. *International Journal of Engineering & Technology*, 7(2.28), 197-199.
- [29] Jusoh, J. A., Man, M., & Bakar, W. A. W. A. (2018). Performance of IF-Postdiffset and R-Eclat Variants in Large Dataset. *International Journal of Engineering & Technology*, 7(4.1), 134-137.
- [30] Agrawal, R., & Srikant, R. (1994, September). Fast algorithms for mining association rules. In *Proc. 20th int. conf. very large data bases, VLDB* (Vol. 1215, pp. 487-499).
- [31] Abu Bakar, W. A. W., Man, M., Abdullah, Z., & Man, M. B. (2021). CRS-iEclat: Implementation of Critical Relative Support in iEclat Model

- for Rare Pattern Mining. *International Journal of Advanced Computer Science and Applications*.
- [32] Rahman, A., Ezeife, C. I., & Aggarwal, A. K. (2012). Wifi miner: An online apriori-infrequent based wireless intrusion detection system. *Knowledge Discovery from Sensor Data (Sensor-KDD 2008)*, 76.
- [33] D. J. Haglin and A. M. Manning, "On Minimal Infrequent Itemset Mining," Proceedings of the International Conference on Data Mining, DMIN'07, CSREA Press, pp. 141-147, 2007.
- [34] A. Gupta, et al., "Minimally Infrequent Itemset Mining Using Pattern-Growth Paradigm and Residual Trees," CoRR abs/1207.4958, 2012.
- [35] S. Tsang, et al., "RP-tree: Rare Pattern Tree Mining," DaWaK, Lecture Notes in Computer Science, Alfredo Cuzzocrea and Umeshwar Dayal (Eds.), Springer, Berlin, vol. 6862, pp. 277-288, 2011.
- [36] A. Aziz, N. H. Ismail, F. Ahmad, Z. Abidin, K. G. Badak, and M. Candidate, "MINING STUDENTS' ACADEMIC PERFORMANCE," undefined, 2013.
- [37] Bakar, W. A. W. A., Man, M., Man, M., & Abdullah, Z. (2020). I-Eclat: Performance enhancement of Eclat via incremental approach in frequent itemset mining. *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, 18(1), 562-570.
- [38] Fournier-Viger, P., Lin, J. C. W., Kiran, R. U., Koh, Y. S., & Thomas, R. (2017). A survey of sequential pattern mining. *Data Science and Pattern Recognition*, 1(1), 54-77.
- [39] Shrivastava, S., & Johari, P. K. (2016, May). Analysis on high utility infrequent ItemSets mining over transactional database. In *2016 IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT)* (pp. 897-902). IEEE.
- [40] Keste, P. A., & Shaikh, N. F. (2016, March). Improved approach for infrequent weighted itemsets in data mining. In *2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT)* (pp. 2663-2667). IEEE.
- [41] Cui, Y., Gan, W., Lin, H., & Zheng, W. (2022). FRI-miner: fuzzy rare itemset mining. *Applied Intelligence*, 52(3), 3387-3402.
- [42] Lu, Y., Richter, F., & Seidl, T. (2020). Efficient infrequent pattern mining using negative itemset tree. In *Complex Pattern Mining* (pp. 1-16). Springer, Cham.
- [43] Man, M., Julaily, A. J., Saany, S. I. A., Bakar, W. A. W. A., & Ibrahim, M. H. (2019). Analysis study on R-Eclat algorithm in infrequent itemsets mining. *International Journal of Electrical and Computer Engineering*, 9(6), 5446.

# A Fully Immersive Virtual Reality Cycling Training (vProCycle) and its Findings

Imran Bin Mahalil<sup>1</sup>, Azmi Bin Mohd Yusof<sup>2</sup>, Nazrita Binti Ibrahim<sup>3</sup>,  
Eze Manzura Binti Mohd Mahidin<sup>4</sup>, Ng Hui Hwa<sup>5</sup>

College of Computing & Informatics, Universiti Tenaga Nasional,  
Putrajaya Campus, Jalan IKRAM-UNITEN, 43000 Kajang, Selangor, Malaysia<sup>1, 2, 3, 4</sup>  
National Sport Institute of Malaysia, Kuala Lumpur, Malaysia<sup>5</sup>

**Abstract**—Virtual reality (VR) technology is popularly applied in various sports training such as cycling, rowing, soccer, tennis and many more. In VR cycling, however, cyclists are not able to fully immerse themselves during the training due to the hardware and applications limitations required in the setup. In order to be fully immersed during the training, cyclists need to have similar effects to an outdoor training where they will experience cycling resistance, temperature effect, altitude, visual, and audio. For this reason, dedicated stimulus effectors or hardware are required to create these expected effects. On cycling resistance, a realistic cycling experience can be simulated by using a special device that simulates a resistance to the back wheel when cycling uphill in the VR simulation. In addition, the back wheel resistance would need to match the view displayed while paddling on an elevation slope. For higher immersion purposes, and the effect of temperature must be created that matches with the view visible in the display. For example, while the cyclist is on top of a virtual mountain, the cyclist would want to feel the effects of high altitude and low temperature. These stimulus effectors affect the realism experience while cycling in the VR simulation training. In the authors' previous papers, the setup using a combination of stimulus effectors including uphill elevation climb, altitude, temperature, interaction, visual, and audio were integrated into a product called vProCycle. The study tested on vProCycle was conducted with an assumption that virtual reality can enhance the experience of physical cycling training. The objective of this study is to determine whether or not vProCycle may improve cyclists' performance. This paper will discuss in detail the findings from data gathered during the experiment using vProCycle. More specifically, the findings are focused on the speed and the heart rate beats per minute which determine their performance improvement.

**Keywords**—Virtual Reality (VR); presence level; technology acceptance; cycling performance; VR cycling training; vProCycle

## I. INTRODUCTION

In this research, several literatures that focus on VR-based cycling training were reviewed in order to identify the current status of the effectors. This paper covers several topics including LR, system setup, experiment, conclusion and future works. According to Oxford Dictionary 2021, the biological meaning of effectors refers to a substance that carries out a response to the stimulus. Furthermore, stimulus refers to a substance or event that evokes a specific functional reaction to the human body. From the literature review, it is found that stimulus-effectors such as snow, water, and wind-effect can be applied in an event that evokes a specific reaction to the body

[1]. In a VR-based training, several stimulus-effectors can be applied to replicate an actual life experience. According to Bohil Corey, "VR technology creates a simulation across the effector arrays and is attached to devices to create an immersive virtual reality system" [1, 2]. Until now, the commonly used VR effectors among cyclists are visual, audio, and interaction [2, 3]. It is suggested by a few researchers that adding more effectors on top of the regular effectors (i.e. audio, visual and interaction) will generate a more immersive experience [3, 4, 5]. In the area of sports, researchers have used different types of combination effectors to improve cyclists' performance [4]. In relation to cycling performance training, it is suggested by the researchers that the following six effectors are utilised: altitude, uphill elevation climb, realistic visuals and audio, realistic interaction, and temperature. As discussed in the authors' previous study, these effectors are obtained from the survey conducted [6, 7].

This research focuses on Virtual Reality (VR) performance cycling training setups using a combination of stimulus effectors as explained in the authors' previous paper [5, 6]. From the author's previous study, vProCycle setup integrated a selected combination of effective stimulus effectors to be used for VR cycling training. The previous author's papers were focused on identifying which stimulus effector is most effective to be selected and integrated into vProCycle [5, 6]. vProCycle uses a combination of stimulus effectors: uphill elevation climb, altitude, temperature, interaction, visual, and audio. In the author's previous study, the technology acceptance and presence towards the vProCycle were analysed during the pilot study [7]. From this pilot study, Imran concluded that the vProCycle's participants have given a high score on the technology acceptance and perception of presence [7]. On the other hand, this paper focuses on the cyclists' heart-rate level measured in Beat Per Minute (BPM), and speed measured in Kilometres Per Hour (KMPH). The data collected can then be used to measure the cyclists' performance while undergoing physical training.

The key contribution of this paper can be divided into three. The first contribution of this paper is identifying the stimulus effectors that can be used for the vProCycle setup. The second contribution is the finding on the correlation between presence level and technology acceptance of the vProCycle setup. The third contribution is related with performance level where it is found that the new setup can be an effective alternative for cycling training. Based on these findings, it is logical to assume

that vProCycle can be used for producing a well-prepared cyclist.

## II. LITERATURE REVIEW

Training using vProCycle has several benefits. As mentioned earlier, vProCycle consists of uphill elevation climb, altitude, temperature, interaction, visual and audio. The first benefit is generated by uphill elevation climb. The positive effect of uphill elevation climb is supported by McIlroy and Al-Kefagy [8, 9]. They have found that the cyclists benefit from the training when they are able to view an uphill elevation climb displayed on the head mounted device. McIlroy and Al-Kefagy also claimed that the interactive view triggers the cyclists to exert more power to paddle forward in the simulation [8, 9]. When the cycling route changes from a steep elevation slope to a flat surface, the elevation would decline and less power is needed to paddle forward. As a result, cyclists will reduce the power required [9, 10, 11]. In the study conducted by Al-Kefagy [9], an urban city environment was used in his VR application E-bike. In his study, the back wheel resistance machine improved the cyclists' safety awareness, their balances, and increased their immersion level significantly.

The second stimulus effector that benefits the cyclist is called altitude. Training in a high altitude with less oxygen can prepare the cyclist for endurance to strengthen the lung capacity [10, 11, 12]. In a study conducted by Hoeg [10], altitude together with virtual reality technology were tested using the cycling track of the Tour de France. The participants were two elite cyclists training for competition where it was found that their performance had significantly improved. The experiment was conducted at 2000 feet above sea level with 20 percent less oxygen molecules in the air as compared to on the sea level [10]. Training at higher altitude would make cyclists to feel more difficult to cycle and get tired more quickly. When a cyclist trains at a higher altitude, the red blood cell delivers less oxygen to the muscles because of the lesser oxygen level in the air. The oxygen is used to produce energy which helps the muscles to move and perform activity. When training multiple times in high altitude, after the third week, the red blood cells will begin to produce more in the cyclist's body [11, 12]. The basic cycling training using altitude begins at 2000 feet above sea level for thirty minutes, one session a week for three weeks. Within the first 7-10 days, there is little to no difference in the number of red blood cells produced in the body. For the next two weeks, the human body produces more red blood cells. In the third week the human body begins to produce a significant amount of red blood cells that affect the cyclist to perform better [11, 12, 13].

The third stimulus effector is temperature, where it is set according to the altitude height. For example, if the height is set at 2000 metres above sea level, the temperature will be set to 20 degrees Celsius. Lower temperature benefits the cyclists to adapt to various body conditions which helps in physical endurance [13, 14, 15].

The last stimulus effectors are called visual and audio. Visual and audio benefit the cyclist by being able to view and hear the virtual environment. Visual and audio can benefit the cyclist when other stimulus effectors are working

synchronously [16, 17]. For example, if the visual displayed to the cyclist is on top of a mountain, then the effect experienced by the cyclist has to match with the view [18, 19, 20, 21].

In order to create an immersive experience simulation setup, integration of various effective stimulus effectors are required [5, 6]. Using a particular set of stimulus effectors, cyclists can adapt to the uphill terrain to improve performance. Adaptation to the uphill terrain cycling training can also be achieved by the familiarisation to the environment [22, 23]. In the study conducted by Mehdi [24], the seven cyclists were familiarised within three fixed-duration sessions of 12 minutes training. Cyclists were able to determine how much power output is required to reach the maximum speed required to reach the destination. In order to determine the effectiveness of the familiarisation to the training terrain, a better performance of speed and heart rate level from the cyclists need to be improved over the three sessions [25, 26, 27]. In another study, Darvish conducted a VR experiment of familiarisation sessions set for a continuous eight minutes where the speed was held for the final three minute[28]. In his experiment, the speed and heart rate (HR) of all 18 cyclists had improved in two sessions [28]. This indicates that the performance can be improved when cyclists are familiar with the environment and more immersive simulation experience would provide a better realism effect for the cyclists.

According to Mascaret an effective VR-based setups used for sport physical training requires a high level of technology acceptance and perception of presence immersion fidelity [29]. When the technology acceptance level is high together with the sense of presence, VR-based training will benefit athletes in improving their sport performance [29].

Table I shows the list of VR-based bicycle setup and its effectors. In this table, four of the previous research setups including vProCycle are shown where the differences between each setup in terms of effectors applied are listed. These four setups provide distinguished experiences to the cyclists.

TABLE I. LIST OF VR SETUP AND EFFECTORS

No.	List of VR bicycle setup and effectors	
	VR-based bicycle setup	Effectors
1	vProCycle	Altitude, uphill elevation climb, temperature, audio and high realistic visuals
2	Al-Kefagy [9]	Uphill elevation climb, audio and visuals
3	Hoeg [10]	Altitude, temperature, uphill elevation climb, audio and less realistic visuals
4	Wu [16]	Uphill elevation climb, audio and visuals

Table I shows four distinctive VR-based cycling training setups using different types of effectors. In the study conducted by Wu [16], he integrated uphill elevation climb, audio and visual into his VR-based bicycle setup. The visuals of the track used in Wu's experiment was based on the 360 recording video. Wu's [16] result shows that the cyclists were able to improve their speed performance using the setup. In another research conducted by Hoeg [10], he added another effector that simulates the altitude effects of a virtual mountain environment. However, in his research, the effector to generate

temperature was not integrated. It is expected that without this effector, the cyclist's would be less immersed as compared to the real world. In the experiment conducted by Al-Kefagy [9], the setup was similar to Wu's, except for the track displayed inside the HMD which was based on the 3D visuals environment created by the VR engine. It is suggested that with a 360 recording video the cyclists would be more immersed.

As discussed above, using a limited set of effectors may improve cyclists' performance. It is also anticipated that cyclists' perception of presence will improve when more effectors are integrated into the system [7].

### III. SYSTEM SETUP

This paper is a continuation from the authors' previous research [5, 6, 7], where the vProCycle used during the experiment is discussed in detail.

Table II shows the list of effectors and technology providers used in vProCycle. The six effectors are matched to a specific technology that provides a distinguished experience for the user.

The first effector, as shown in Table II is audio and visual. In this setup, HMD is used as the technology provider. There are many brand names of HMD available in the market such as Oculus and HTC VIVE. In this research, the Oculus Quest 2 unit was used to produce a full 360 degree viewing angle with an immersive audio effects. The second effector is called altitude effect. This effect is gained by using a chamber room, as seen in Fig. 1.

Fig. 1 shows a chamber room and the control panel that simulate altitude, temperature, oxygen level and humidity level. Fig. 1(a) shows the chamber room from the outside view and Fig. 1(b) shows the control panel that adjusts the settings for the simulation. The objective of using a chamber room is to create a high fidelity of realism to the VR simulation cycling training.

TABLE II. LIST OF EFFECTORS AND TECHNOLOGY PROVIDERS

No.	List of effectors and technology provider	
	Effectors	Technology provider
1	Audio and visuals	HMD
2	Altitude	Altitude chamber room
3	Uphill elevation climb	Backwheel resistance machine
4	Temperature	Altitude chamber room
5	Paddling interaction	Bluetooth device

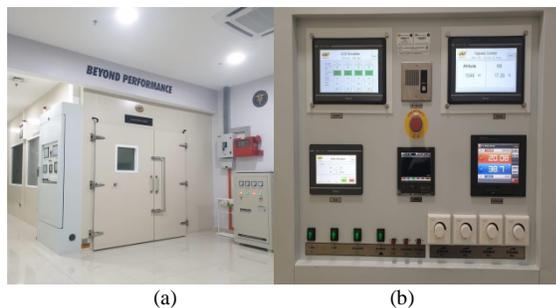


Fig. 1. Chamber room.

The third effector is called uphill elevation climb with the technology provider of a back wheel machine that grips the back wheel of the bicycle to provide a resistance effect. There are many brand names that provide back wheel machines such as Wahoo, Magene, Kinetic, and etc. In this research, Kinetic was used to provide an immersive effect of uphill elevation climb.

The fourth effector is called temperature. The temperature effect is simulated inside a chamber room through the control panel, as seen in Fig. 1(a) and (b).

The last effector is called paddling interaction. The paddling interaction is determined by a bluetooth device that affects the movement inside the virtual world. This bluetooth device is attached onto the bicycle paddle which calculates the Revolutions Per Minute (RPM) in real time. The bicycle in the virtual world would move forward based on the calculated RPM. There are many bluetooth devices available such as Magene, Vzfit, and Wahoo. In this research Magene was used to provide a realistic paddling interaction effect in the virtual simulation. Note that each brands requires their own applications downloaded and controlled using a handphone. Vzfit technology is the major component applied in vProCycle. Vzfit technology consists of the HMD application that connected with backwheel machine using a bluetooth device. Inside the chamber room, the altitude of the height is also set according to the virtual environment seen by the cyclist.

Many literature encourage the use of more distinctive stimulus effectors for a better VR application out-put. The individual stimulus effectors have already been investigated. However, a comprehensive study on a VR application that consists of several stimulus effectors have not been maximised used in the text, even after they have been defined in the abstract. Abbreviations such as IEEE, SI, MKS, CGS, sc, dc, and rms do not have to be defined. Do not use abbreviations in the title or heads unless they are unavoidable.

Fig. 2 shows the setup of a VR cycling training system with the combination of uphill elevation climb, altitude, temperature, interaction, visual, and audio. This VR training system is named as "vProCycle system" with the intellectual property copyright number 2022/CR/15.71/OP.

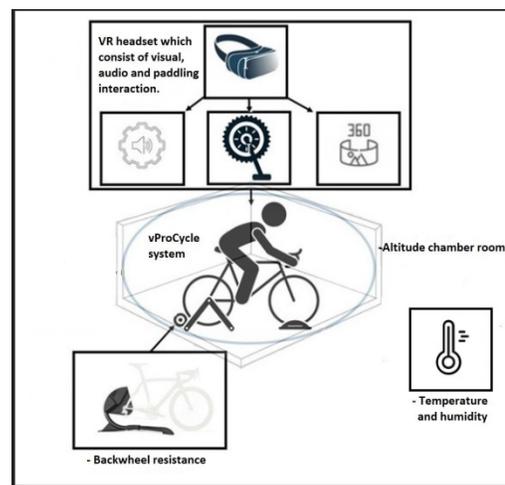


Fig. 2. Setup of the vProCycle

### A. Hardware and Technology Requirement

This experiment was conducted at The National Sport Institute Malaysia, involving 10 volunteers with different background skills. The volunteers were two elite cyclists and eight athletes from a university football team. During the experiment, the cyclists' speed in the virtual environment and the heart rate level (BPM) were collected. In this subsection, the cyclists' requirements and materials used to conduct the experiment are explained in detail below. The figure below shows the setup design during the experiment.



Fig. 3. The vProCycle system setup.



Fig. 4. The vProCycle system setup side view.

The vProCycle system experiment was conducted at the National Sport Institute Malaysia hypoxic chamber as seen in Fig. 3. The experiment setup includes a back-wheel machine, Bluetooth sensor on the bicycle paddle, VR HMD, and heart rate device. This setup was similar to the pilot study with the exception of a heart rate device that was strapped on the chest of the cyclists. The location of the scene in the HMD was L'Étape du Tour, France. Before undergoing a training session, a consent form was signed by the cyclists. The experiment was conducted in three separate sessions with an interval of one week apart, each session was a continuous, non-stop session for 30 minutes. All the cyclists were briefed on the warnings of training in the hypoxic chamber. The temperature and carbon dioxide level was different than on sea level which may cause vomiting, drowsiness, or headaches. Cycling training in

altitude for 30 minutes benefits long term endurance of at least three sessions in an interval of one week [8]. Note that all participants undergone training at 2000 metres above sea level (m.a.s.l), with 20 percent less oxygen as compared to sea level altitude and a temperature of 20 degrees. Fig. 4 shows the side view of the setup in the chamber room using vProCycle; different road bicycles can be mounted.

### B. Experiment Participant Requirement

The selection of the participants were based on the following criteria:

- Age: An individual aged 20 years to 50 years.
- Experience: Have participated in a local or international cycling competition or event.
- Role: Have represented a sport organisation, club or school.
- Type of training: Have trained using a stationary bicycle.
- Demography: No restriction.
- Gender: No restriction.
- Weight: 50kgs-80kgs.
- Height: 1.6m-1.9m.

In this research, ten cyclists that met the above criteria have participated in the experiment. The first criteria is the participants' age. The age is appropriate for cycling performance training [8-13]. The second criteria is the experience of the cyclists. These experiences enable the cyclist to safely undertake cycling training. The third criteria is on the role of the cyclist. The role refers to the types of organisation or entities that they have represented. The fourth criteria is on the type of training they have participated in order to prepare themselves for the upcoming events. As for the fifth, sixth; demography and gender, there is no restriction applied. The seventh and eighth; weight and height, the restrictions are as stated above. The weight and height restrictions are due to the hardware's limitation such as the height and the capacity of the bicycle.

### C. Experimental Material

The cyclists were required to cycle for thirty minutes in a virtual environment based on a realistic view of L'Étape du Tour, France. The distance of the training track of L'Étape du Tour, France was 13 KM in which the cyclists had to cycle as far as they can within the given time duration set. This location was set with many uphill elevation climbs and down hills. Based on the literature review, familiarization of terrain requires a minimum of three sessions and at least eight continuous minutes with an improved performance [26, 27, 28]. In this research, the technology acceptance questions were derived from the original TAM developed by Davis, 1989 [7]. Four questions for each of the TAM's factors were adopted. The presence questionnaire were adopted from the Witmer and Singer Presence Questionnaire. Example of the questions used are "When you bicycle forward, did the feeling of moving forward in the virtual world seem realistic?" and "When you

paddled the bicycle, did you feel the movement speed as realistic according to the change of view?" Both questionnaire (TAM and presence) were given after the cyclists had undergone the training session using vProCycle [7].

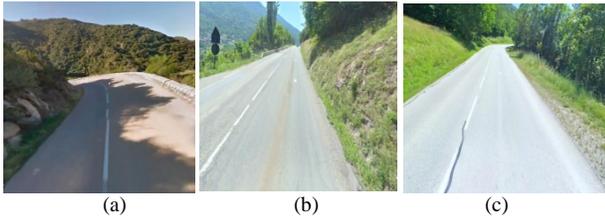


Fig. 5. Three views of L'Étape du Tour, France used during the experiment.

Figure 5 shows three images of the view seen in the HMD at L'Étape du Tour, France. Fig. 5(a) shows the road surface in between one to five degrees of elevation uphill slope curving to the steep left angle. Fig. 5(b) shows a steep elevation slope of 10 degrees. Fig. 5(b) also shows the view of climbing a steep uphill on the road side going straight. Fig. 5(c) shows a downhill slope of zero resistance going downhill curving to the right side. The familiarisation to uphill adaptation from cyclist in altitude requires a minimum of 3 consecutive sessions of an interval of 1 week in between each session [5, 6]. This is due to the body producing more red blood cells in the third week of training in altitude [10, 11, 12].

#### D. Findings

The experiment focuses on two different findings. The first finding is on the average speed of the ten cyclists using vProCycle during performance training. The second finding focuses on the average heart rate level based on the BPM. The two main data collected indicate the cyclists' performance which was measured by the average speed, the heart-rate level and the correlation between the two.

Fig. 6 shows the cyclists' average speed in kilometres per hour while using vProCycle in 3 sessions. P1 to P10 indicates participant 1 to participant 10, respectively. Speed is the distance divided by time. Distance is measured in kilometres and time is measured in minutes. Heart-rate level is measured by the amount of beats per minute (BPM) that the heart pumps.

In session 1, the lowest speed was retrieved from participant 3, 4, 5, and 6 at 10 KMPH, whereas the highest speed was reported at 16.7 KMPH by participant 10 and 12.7 KMPH by participant 9. In session 2, the lowest speed was recorded by participant 5 at 10.8 KMPH, while participant with the highest speed was participant 10 at 17 KMPH. In session 3, the lowest speed was recorded by participant 3 and 5 with both recorded an average speed at 12 KMPH, while the highest speed was recorded at 17 by participant 10. Each participant had improved their performance by increasing their speed from session 1, 2, and 3, subsequently. Speed is measured as the ratio of distance to the time in which the distance was covered, and the time period given for each session was 30mins. From the average speed, it could identify the average Kilometres Per Hour (KMPH), and how many kilometres (KM) they have reached by the end of each session. By capturing the data of KMPH and the KM from each session, analysis can be made to identify whether the participants have improved their speed

performance or not. Below is the analysis of each individual participant. In this analysis, the improvement from one to two KM was considered major while less than one was considered minor [10].

In relation to the KMPH, the cyclists' KMPH were lower in the first session and gradually increased during the consecutive sessions. It is logical that the cyclists' KMPH in the early session was lower due to the unfamiliar training conditions including altitude. It is also found that the changes in terms of KMPH between session 1 and session 2 was much higher as compared to session 2 and session 3.

As a summary, it can be seen that all participants' average speed increased throughout the sessions. All cyclists' speed performance improved when using vProCycle. The following is the findings on the participants' heart rate level based on BPM.

Fig. 7 shows the average heart rate based on the amount of beat per minute (BPM) with 90% of the cyclists' average heart rate (BPM) in session 1 was above 150. The lowest score reported in session 1 was 140 BPM and the highest was 172. In session 2, all cyclists were below 162 BPM whereas in session 1 four cyclists were above 170. The highest BPM score in session 2 was reported at 161 and the lowest was at 130. This would indicate that as compared to session 1, the heart rate BPM for all cyclists had dropped in session 2. This might be due to the fact that the cyclists may have not yet adapted to the new setup and not comfortable in the first session as compared to the second session within a one week interval.

In relation to the BPM, the cyclists who are not familiar with training at 2000 metres above sea level may find discomfort during the first session [24, 25, 26]. Due to the unfamiliar conditions of the altitude and uphill elevations, the cyclists' BPM would be higher in the first session compared to session 2 and session 3. The BPM indicates the endurance level of the cyclist when undergoing training. An increase in BPM usually occurs when the cyclists are required to exert more power when increasing their speed or climbing uphill [24, 25, 26].

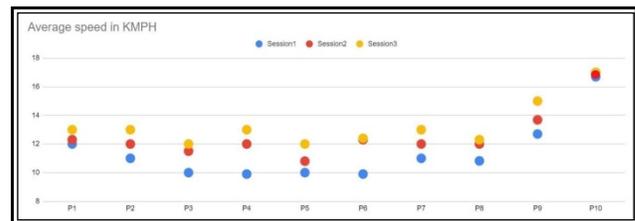


Fig. 6. Participants' average speed in kilometres per hour (KMPH)



Fig. 7. Cyclists' average heart rate in Beat Per Minute (BPM)

As a summary, all participants' BPM decreased throughout the sessions. It shows that their heart rate level improved when using vProCycle. Below are the findings on the participants' average heart rate level correlation to the average speed.

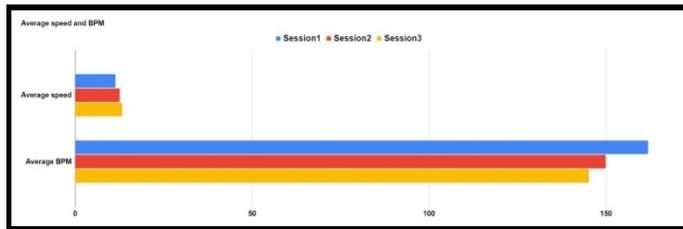


Fig. 8. Cyclists' average speed (KMPH) and heart rate in Beat Per Minute (BPM).

Heart rate level improved when using vProCycle. Below are the findings on the participants' average heart rate level correlation to the average speed.

Fig. 8 shows the cyclists' average speed on the top and the average BPM at the bottom. When the speed is high and the heart rate level BPM is low, this indicates that the cyclist is at a better cycling performance. The average speed of the cyclists in session 1 was reported at 11.402 KMPH, session 2 at 12.559 KMPH, and session 3 at 13.27 KMPH. The average speed in session 1 was at 11.402 KMPH as compared to in session 3 at 13.27 with the difference of 1.9 KMPH. This would indicate that the average speed of the cyclists between session 1 and 3 in this experiment would have improved by 1.9KM. The difference between the average speed of session 1 and 2 was 1.157 KMPH, which indicates that between session 1 and 2, the cyclists improved by 1.157 KMPH on average. For session 2 and 3, the average speed improved by 0.7 KMPH. There was a gradual increase of average speed from session 1, 2, and 3.

As shown in Fig. 8, the improvement in cyclists' performance was also found when analysing the negative correlation between the average speed and the average heart rate level. A negative correlation between the average speed and the average BPM means that while the speed is increasing, the BPM is also decreased.

Session 1 average BPM was at 161.9 BPM as compared to session 3 at 145.2 BPM, with the difference of 16.7 BPM. This would indicate that the cyclists have improved their BPM by 16.7 between session 1 and session 3. The difference between the BPM of session 1 and 2 was 12 BPM, which indicates that between session 1 and 2 the cyclists improved their BPM by 12. For session 2 and 3, BPM was improved by 4.7 BPM. There was a gradual decrease of BPM from session 1, 2 and 3. The lower the heart rate over session would mean that the cyclist is improving.

There is a correlation between speed KMPH and heart rate BPM, in which the higher the speed KMPH across the sessions would decrease the heart rate BPM. The reason for this is because as the cyclists train more frequently, their physical body become more adapted to the higher altitude, thus creating more red blood cells and is able to endure the harsh conditions of this cycling performance training.

As discussed above, all the participants' speed increased throughout the three sessions while their heart rate beat dropped. This indicates that while using vProCycle, participants' performances have improved as there is a negative correlation between the average speed and BPM from all cyclists. As discussed in the author's previous paper [7], the cyclists' were required to answer a questionnaire on TAM and perception of presence. In addition to TAM and perception of presence, an open interview was also conducted. From the interview, it was found that during the first session, all ten cyclists were not familiar with the altitude and uphill elevation climb which affected their performance. During the second session, all the cyclists stated that they had adapted to the training better. In the final session, the cyclists highlighted that they feel more confident resulting in better performance. Based on the average speed and heart rate level, all cyclists showed that they have improved in performance during the final session when using vProCycle.

#### E. Experimental Outcome

From this research, the outcome shows that cyclists' speed performance and heart rate have improved when cycling in the simulated VR terrain using vProCycle. This outcome also suggests that the cyclists' technology acceptance and perception of presence level as explained in the previous paper [5, 6] may have positively impacted the cyclists' performances. In the previous paper, it was found that cyclists gave high rating to both technology acceptance and perception of presence. The findings on the speed and heart rate level also suggest that with a high level of technology acceptance and perception of presence, positive correlations can be seen where cyclists' performance have improved when using vProCycle.

#### IV. CONCLUSION AND FUTURE WORK

In conclusion, this study shows how vProCycle may offer new possibilities in cycling training by combining various stimulus effectors. It is highlighted that distinctive stimulus effectors do greatly influence the performance level of participating cyclists when familiarized with the simulated terrain. This performance level is experienced by a realistic simulation integrated with the system. This simulation creates an experience as though the cyclist is cycling in the distinctive actual real environment. This high level of cyclists' performance thus generates a gradual improvement by the speed and heart rate level throughout the sessions.

It is found that the cyclist's performance has improved based on the real time data collected from the devices (RPM and heart beat rate) attached to the cyclist during the training. A negative correlation between the RPM and heart rate shows that there is an improvement in performance over three sessions.

Cyclists have positively accepted the vProCycle based on technology acceptance and perception of presence level questionnaires. Findings show a positive correlation between RPM and heart rate level with technology acceptance and perception of presence.

This study contributes to the fact that using highlighted distinctive stimulus effectors do greatly influence the performance level of participating cyclists. VR as a tool can be

used to simulate an effective cycling training destination when integration is set distinctively. For future work, a motion platform will be integrated with the vProCycle in order to simulate a tilting effect. This tilting effect occurs when cyclists are turning their direction based on the VR simulated road condition.

## V. INTELLECTUAL PROPERTY

The findings from this paper were conducted using vProCycle. vProCycle has been copyrighted (2022/CR/I5.71/OP).

## ACKNOWLEDGMENT

This work is supported by the Universiti Tenaga Nasional BOLD research grant. This work is supported by the Universiti Tenaga Nasional BOLD research grant.

## REFERENCES

- [1] Bohil CJ, Alicea B, Biocca FA. Virtual reality in neuroscience research and therapy. *Nature reviews neuroscience*. 2011 Dec;12(12):752-62. J. Clerk Maxwell, *A Treatise on Electricity and Magnetism*, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.
- [2] McIlroy B, Passfield L, Holmberg HC, Sperlich B. Virtual Training of Endurance Cycling—A Summary of Strengths, Weaknesses, Opportunities and Threats. *Frontiers in Sports and Active Living*. 2021 Mar 4;3:31.
- [3] Li W, Xie B, Zhang Y, Meiss W, Huang H, Yu LF. Exertion-aware path generation. *ACM Trans. Graph.*. 2020 Jul 1;39(4):115.
- [4] Yang YS, Koontz AM, Hsiao YH, Pan CT, Chang JJ. Assessment of Wheelchair Propulsion Performance in an Immersive Virtual Reality Simulator. *International Journal of Environmental Research and Public Health*. 2021 Jan;18(15):8016.
- [5] Mahalil I, Yusof AM, Ibrahim N. A literature review on the usage of Technology Acceptance Model for analysing a virtual reality's cycling sport applications with enhanced realism fidelity. In 2020 8th International Conference on Information Technology and Multimedia (ICIMU) 2020 Aug 24 (pp. 237-242). IEEE.
- [6] Mahalil I, Yusof AM, Ibrahim N. A literature review on the effects of 6-Dimensional virtual reality's sport applications toward higher presence. In 2020 8th International Conference on Information Technology and Multimedia (ICIMU) 2020 Aug 24 (pp. 277-282). IEEE.
- [7] Mahalil I, Yusof AM, Ibrahim N, Ng Hui W, Eze M. Users' Acceptance and Sense of Presence towards VR Application with Stimulus Effectors on a Stationary Bicycle for Physical Training. *International Journal of Advanced Computer Science and Applications (IJACSA)*, Volume 13 Issue 6, page 56-64, 2022.
- [8] McIlroy B, Passfield L, Holmberg HC, Sperlich B. Virtual Training of Endurance Cycling—A Summary of Strengths, Weaknesses, Opportunities and Threats. *Frontiers in Sports and Active Living*. 2021 Mar 4;3:31.
- [9] Al-Kefagy M, Pokrajac S. E-bike Simulator—a virtual reality application that evaluates user interfaces in an urban traffic environment. 2021.
- [10] Hoeg ER, Bruun-Pedersen JR, Cheary S, Andersen LK, Paisa R, Serafin S, Lange B. Buddy biking: a user study on social collaboration in a virtual reality exergame for rehabilitation. *Virtual Reality*. 2021 Jul 27:1-8.
- [11] Townsend NE, Gore CJ, Ebert TR, Martin DT, Hahn AG, Chow CM. Ventilatory acclimatisation is beneficial for high-intensity exercise at altitude in elite cyclists. *European journal of sport science*. 2016 Nov 16;16(8):895-902.
- [12] Fan JL, Bourdillon N, Meyer P, Kayser B. Oral Nitrate supplementation differentially modulates cerebral artery blood velocity and prefrontal tissue oxygenation during 15 km time-trial cycling in normoxia but not in hypoxia. *Frontiers in physiology*. 2018 Jul 16;9:869.
- [13] Sorensen A, Aune TK, Rangul V, Dalen T. The validity of functional threshold power and maximal oxygen uptake for cycling performance in moderately trained cyclists. *Sports*. 2019 Oct;7(10):217.
- [14] Karetnikov A. Application of data-driven analytics on sport data from a professional bicycle racing team. Eindhoven University of Technology, The Netherlands. 2019.
- [15] Cui T, Yang Y, Guo Y. Evaluation of Height and Speed Effects on the Comfort of VR Motion Picture Display. In 2021 International Conference on Culture-oriented Science & Technology (ICCST) 2021 Nov 18 (pp. 426-430). IEEE.
- [16] Wu TF, Tsai PS, Hu NT, Chen JY, Huang YS. Reality Simulation for Bike Training Devices with Touch Panel. *Sensors and Materials*. 2018 Jan 1;30(3):609-20.
- [17] Westmattmann D, Grotenhermen JG, Sprenger M, Rand W, Schewe G. Apart we ride together: The motivations behind users of mixed-reality sports. *Journal of Business Research*. 2021 Sep 1;134:316-28.
- [18] Matvienko A, Müller F, Zickler M, Gasche LA, Abels J, Steinert T, Mühlhäuser M. Reducing Virtual Reality Sickness for Cyclists in VR Bicycle Simulators. In CHI Conference on Human Factors in Computing Systems 2022 Apr 29 (pp. 1-14).
- [19] Nazemi M, van Eggermond MA, Erath A, Schaffner D, Joos M, Axhausen KW. Studying bicyclists' perceived level of safety using a bicycle simulator combined with immersive virtual reality. *Accident Analysis & Prevention*. 2021 Mar 1;151:105943.
- [20] Shoman MM, Imine H. Bicycle Simulator Improvement and Validation. *IEEE Access*. 2021 Apr 5;9:55063-76.
- [21] Cui T, Yang Y, Guo Y. Evaluation of Height and Speed Effects on the Comfort of VR Motion Picture Display. In 2021 International Conference on Culture-oriented Science & Technology (ICCST) 2021 Nov 18 (pp. 426-430). IEEE.
- [22] Mackey J, Horner K. What is known about the FTP20 test related to cycling? A scoping review. *Journal of Sports Sciences*. 2021 Dec 2;39(23):2735-45.
- [23] Davies MJ, Clark B, Welvaert M, Skorski S, Garvican-Lewis LA, Saunders P, Thompson KG. Effect of environmental and feedback interventions on pacing profiles in cycling: a meta-analysis. *Frontiers in Physiology*. 2016 Dec 5;7:591.
- [24] Kordi M, Fullerton C, Passfield L, Parker Simpson L. Influence of upright versus time trial cycling position on determination of critical power and  $W'$  in trained cyclists. *European Journal of Sport Science*. 2019 Feb 7;19(2):192-8.
- [25] Treweek N, Neumann DL, Hamilton K. Effect of affective feedback and competitiveness on performance and the psychological experience of exercise within a virtual reality environment. *PloS one*. 2022 Jun 8;17(6):e0268460.
- [26] Wender CL, Tomporowski PD, Ahn SJ, O'Connor PJ. Virtual reality-based distraction on pain, performance, and anxiety during and after moderate-vigorous intensity cycling. *Physiology & Behavior*. 2022 Jun 1;250:113779.
- [27] Guo X, Robartes E, Angulo A, Chen TD, Heydarian A. Benchmarking the use of immersive virtual bike simulators for understanding cyclist behaviors. In *Computing in Civil Engineering 2021* 2021 (pp. 1319-1326).
- [28] Darvish S, McNulty M, Pon J, Tallarida H, Moody J, Conant S, Thorp DB. The Effect of Visual Flow on Cycling in a Virtual Environment. In *International Journal of Exercise Science: Conference Proceedings 2020* (Vol. 8, No. 8, p. 47).
- [29] Mascaret N, Montagne G, Devriese-Sence A, Vu A, Kulpa R. Acceptance by athletes of a virtual reality head-mounted display intended to enhance sport performance. *Psychology of Sport and Exercise*. 2022 Jul 1;61:102201.

# First Responders Space Subdivision Framework for Indoor Navigation

Asep Id Hadiana<sup>1</sup>, Safiza Suhana Kamal Baharin<sup>2</sup>, Zahriah Othman<sup>3</sup>

Center of Advanced Computing Technology (CACT)-Faculty of Information and Communication Technology, Universiti Teknikal Malaysia Melaka, Melaka, Malaysia<sup>1,2,3</sup>  
Department of Informatics, Universitas Jenderal Achmad Yani, Cimahi, Indonesia<sup>1</sup>

**Abstract**—Indoor navigation is crucial, particularly during indoor disasters such as fires. However, current spatial subdivision models struggle to adapt to the dynamic changes that occur in such situations, making it difficult to identify the appropriate navigation space, and thus reducing the accuracy and efficiency of indoor navigation. This study presents a new framework for indoor navigation that is specifically designed for first responders, with a focus on improving their response time and safety during rescue operations in buildings. The framework is an extension of previous research and incorporates the combustibility factor as a critical variable to consider during fire disasters, along with definitions of safe and unsafe areas for first responders. An algorithm was developed to accommodate the framework and was evaluated using Pyrosim and Pathfinder software. The framework calculates walking speed factors that affect the path and walking speed of first responders, enhancing their chances of successful evacuation. The framework captures dynamic changes, such as smoke levels, that may impact the navigation path and walking speed of first responders, which were not accounted for in previous studies. The experimental results demonstrate that the framework can identify suitable navigation paths and safe areas for first responders, leading to successful evacuation in as little as 148 to 239 seconds. The proposed framework represents a significant improvement over previous studies and has the potential to enhance the safety and effectiveness of first responders during emergency situations.

**Keywords**—Space subdivision; indoor navigation; first responders; indoor disaster

## I. INTRODUCTION

Most of human life is spent indoors rather than outdoors. The average time spent indoors was 87%[1]. The increasing concentration of human populations in modern urbanized societies has aggravated the frequency and destruction of natural and artificial disasters. A United Nations study indicates that by 2050, 66% of the world's population is expected to be urban. Hence, buildings are increasingly significant, broad, and complex to accommodate them and guarantee all the requirements for protection and good well-being [2]. Buildings have been the primary focus of recent disasters, resulting in heavy damages and losses for public and emergency managers.

Residential building fires and fire deaths tend to increase yearly [3]. The increasing concentration of human populations in modern urbanized societies has aggravated the frequency and destruction of natural and manmade disasters. First responders are critical in responding to indoor emergencies,

such as those in high-rise buildings. First responders, also called emergency responders, are those in charge of saving lives, protecting property, and keeping the environment safe in the early stages of an accident or disaster[4]. In large or complex buildings, it can be difficult for first responders to navigate through the entire building at once, particularly in an emergency where time is of the essence. Indoor navigation is a critical aspect of emergency response. It allows first responders, such as firefighters and emergency medical technicians, to locate and assist those in need within a building or other indoor space quickly and safely.

To effectively navigate an indoor space, first responders need to locate themselves accurately within the space and determine the most efficient route to their destination. Indoor navigation can help first responders more quickly and efficiently navigate the building, allowing them to focus on responding to emergencies and providing assistance to those in need[5]. Indoor navigation is particularly challenging because buildings can be complex, with multiple floors, rooms, and corridors, making it difficult to determine one's location and direction. This is further complicated by the presence of smoke, low visibility, and other hazards that are often present during an emergency.

In an emergency, first responders must determine the safest route through the building to reach their destination as quickly and safely as possible. This may involve considering the combustibility of the building materials, the presence of hazards or obstacles, and the location of windows or other alternative exits. In large or complex buildings, it can be difficult for first responders to navigate through the entire building at once, particularly in an emergency where time is of the essence.

One approach to improving indoor navigation for first responders is using space subdivision. This technique involves dividing a building or other indoor space into smaller, more manageable units or sections, making it easier to navigate and orient oneself within the space. In addition to improving navigation, space subdivision can also help to improve communication and coordination among first responders. By dividing the space into smaller units, it becomes easier to provide specific and accurate directions and instructions and to communicate the location and status of individuals or resources within the space. The goal of space subdivision is to provide first responders with detailed, real-time information about the layout and occupancy of a building, which can help

them to reach the scene of an incident more quickly and safely.

In emergencies, there may be dynamic changes in the building, such as the spread of fire or smoke. It can be crucial for first responders to continuously update their navigation system to reflect these changes to stay safe and reach their destination as quickly as possible. Space subdivision can be used to identify which areas of the building are safe to navigate at any given time and to provide real-time updates to the navigation system.

In emergencies, first responders must determine the safest route through the building to reach their destination as quickly and safely as possible. Space subdivisions can be used to identify spaces with lower combustibility or windows that can be used as alternative exit routes and to determine the safest route through the building based on these characteristics. Overall, space subdivision can support indoor navigation for first responders in emergencies by allowing them to focus on navigating through smaller, more manageable areas of the building and staying safe.

Subdividing indoor areas is an important part of scene analysis that is used for a wide variety of purposes, including but not limited to navigation and evacuation planning [6]. Space Subdivision plays a vital role in indoor navigation [7]. The division of space in indoor navigation is basically to determine navigable and non-navigable spaces.

To improve the accuracy and efficiency of indoor navigation, current spatial subdivision models need to reflect the indoor context's dynamic changes ultimately. However, they also need help defining precise navigable space [8]. When responding to emergencies, first responders must be aware of the unexpected dynamic changes in indoor space. Hence a subdivision structure of indoor space that adapts to varied indoor spatial features and spatiotemporal dynamics is of considerable importance [9].

Several previous studies have discussed the space subdivision for indoor navigation. However, these studies have yet to reflect the dynamic conditions that often change when a disaster occurs in a building, especially during rescue operations by first responders. For this reason, this study proposes a space subdivision framework for indoor navigation for first responders.

This research aims to provide a new framework for indoor navigation that is specifically designed for first responders in emergency situations. This framework is based on the idea of dividing indoor spaces into smaller subspaces, which can be more easily navigated and searched by first responders. Some potential benefits of this approach are:

1) *Improved efficiency*: The proposed framework can help improve the efficiency of first responders during emergency situations. By providing a more structured and organized approach to navigating indoor spaces, first responders can quickly and efficiently locate individuals who need help or identify potential hazards.

2) *Enhanced safety*: The subdivision of indoor spaces can also help improve the safety of first responders by allowing

them to more easily identify potential hazards, such as structural damage or chemical spills. This can help reduce the risk of injury or harm to both first responders and those they are trying to help.

The main contributions of this research include the development of a novel framework for indoor navigation that takes into account the dynamic changes of the indoor context during fire emergencies, as well as the identification of suitable navigation paths and safe areas for first responders. The framework also incorporates an algorithm that is tested using Pyrosim and Pathfinder software, which calculates the walking speed factor and affects the path and walking speed of first responders. The experimental results demonstrate that the proposed framework can improve response times and increase the chances of successful evacuation for first responders.

The implications of this research are significant, as it can improve the safety and effectiveness of first responders during rescue operations in indoor environments. By accounting for dynamic changes in the indoor context, the framework can identify suitable navigation paths and safe areas for first responders that may not have been considered in previous research. The framework has the potential to become an essential tool for emergency response teams, enhancing their ability to navigate indoor environments during emergencies and improving their chances of success.

This rest of this article is organized as follows: Section II provides a review of related work. Section III outlines the proposed framework for indoor navigation, which involves dividing the space into smaller sections. Section IV goes into further detail about space subdivision and network derivation. Section V explains the algorithm used for path planning. Section VI describes the experiments that were conducted. Section VII presents and discusses the results of the implementation. Finally, Section VIII concludes the research by summarizing the findings and outlining future research directions.

## II. RELATED WORK

Some literature has extensively investigated the subdivision of space for indoor navigation. The approach usually uses well-defined building construction spaces such as rooms and corridors, builds connectivity between these spaces that use their semantics, and applies Poincaré Duality to obtain navigable networks [10]. Several semantic models have been established in this context to help with space identification and develop a navigation network [11]. The generation of an effective navigation network heavily depends on a reasonable subdivision of indoor space [12].

To create navigable subspaces for refined indoor navigation, indoor spatial subdivision, as the primary method of indoor space organization, has been extensively studied. The study [13] identifies four different steps necessary for a successful application for navigation through indoor spaces. The digital acquisition of available spaces (1), the structuring of acquired data (2), formalization of the data to establish relationships between different subspaces (3) and lastly applying the user requirements on the formalized and structured data (4). The subdivision of indoor space forms

subspaces into smaller parts. The research [14] divides indoor spaces into navigable and non-navigable areas by considering human social behaviour at different times and then redivide the navigable space by the Constrained Delaunay Triangulation.

Several methods for partitioning the fundamental indoor space have been developed in order to construct fine-grained indoor navigable space. It is a common practice to use a dedicated strategy (such as visibility graph, Delaunay triangulation, and convex hull cell) or regular gridding (such as square, hexagon, etc.) [15] to divide an entire indoor space in order to derive a 2D navigable network. This can be accomplished in a number of different ways.

A fine-grained and context-aware subdivision framework (FSS) is proposed in [10] to remodel the relevant 3D space for the arrangement of navigable indoor environments by injecting numerous materials. This is accomplished by rearranging the components in a specific order. The FSS framework is hailed as a landmark for its ability to depict indoor human interaction behaviors to date. It is an inspiration for better indoor environment modeling and cognition-based navigation.

The initial formation of the F-Space considers dynamic human activities; however, their significance in the fine-grained spatial subdivision of the indoor environment is overlooked. As a result, the F-Space needs to be fully navigable possible. Within this framework, the authors separated things into three categories according to their mobility: the capacity to change their location (static, semi-mobile, and mobile). After then, the interior space, also known as the environment, was divided into three distinct sections: object space, functional-space, and remaining-space. To be more specific, object space is the non-navigable portion of subspace that is occupied by semi-mobile objects. Functional space, on the other hand, is the subspace portion dedicated to the utilization of semi-mobile objects or the activities of mobile objects and is navigable under certain conditions.

Depending on the application, subdivisions can be done on 2D or 3D levels [16]. As found in [17], there are certain advantages and disadvantages of conducting 2D and 3D subdivisions. The key advantages of 2D subdivision are that all or most calculations may be performed in 2D, the user can retain just places occupied by pedestrians in memory, and triangles are not required for vertical pedestrian placement. On the other hand, if everything is viewed as a single 3D space or pieces of 3D spaces, simulated pedestrians can freely move throughout an area without having to check to see if they have entered another space for pedestrian dynamics purposes.

### III. SPACE SUBDIVISION FRAMEWORK

When conducting emergency rescues, first responders are often only provided with a floor plan map, even though the map already contains many objects in actual conditions, as shown in Fig.1.

However, this floor plan map may not accurately reflect the actual conditions of the space during the emergency situation. The map may have been created prior to any

changes to the space or may not be detailed enough to capture all the objects or obstructions that may be present.

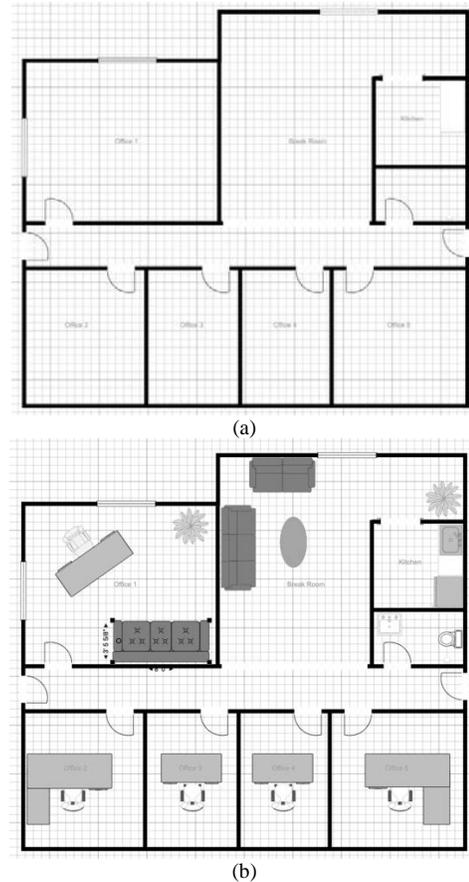


Fig. 1. Comparing the (a) Floor plan to the (b) Actual situation of an indoor space.

In addition, the floor plan map may not provide a clear or accurate representation of the space, making it difficult for first responders to navigate and locate individuals who need assistance. This can be especially challenging in complex indoor environments, such as large buildings or underground tunnels.

The proposed approach from this research, addresses these challenges by dividing the indoor space into smaller subspaces, which can be more easily navigated and searched by first responders. This approach provides a more detailed and accurate representation of the space, which can help first responders better understand the layout and locate individuals who need assistance.

By using a subdivision framework, the first responders can more easily identify potential hazards and obstacles within the space, which can improve their safety and reduce the risk of injury. Additionally, the framework can be integrated with existing technology, such as mapping and tracking systems, to provide real-time information on the location and movement of first responders within the indoor space.

Space subdivision can be a useful tool for indoor navigation for first responders in an indoor emergency situation, as it allows the building to be divided into smaller,

more manageable areas that can be navigated separately. This can be particularly helpful in situations where there are dynamic changes in the building, such as the spread of fire or the presence of smoke, as it allows the first responders to focus on navigating through a smaller area at a time rather than trying to navigate through the entire building at once.

Here are some ways in which space subdivision can help with indoor navigation for first responders in an indoor emergency situation:

1) *Identifying safe navigable spaces:* By dividing the building into smaller spaces, it can be easier to identify which areas are safe to navigate based on factors such as the combustibility of the material and the presence of hazards.

2) *Determining the safest route:* By dividing the building into smaller spaces, it can be easier to determine the safest route through the building based on the characteristics of each space. For example, the first responders might choose to navigate through spaces with lower combustibility or with windows that can be used as alternative exit routes.

3) *Providing real-time updates:* By continuously monitoring the building for dynamic changes and updating the space subdivision accordingly, it can be easier to provide real-time updates to the navigation system to reflect any changes in the safe navigable spaces.

We propose a classification of interior objects based on mobility and safety for first responders as a means of subdividing dynamic indoor settings. This classification would be used by first responders. This classification expands [10] original works.

In order to get a head start on our work, we will first go over the essential definitions of indoor objects and indoor subspaces following the FSS framework from [10]. The FSS framework categorizes indoor objects as either static S-objects (such as indoor fixed structures), mobile M-objects (such as humans), or semi-mobile SM-objects. Static S-objects include indoor fixed structures. The products have been sorted into these categories according to how mobile they are. Independent of the type of building being used, the S, SM, and M-objects provide direction for the space subdivision. They are utilized in defining the geometry, semantics, and topology of the subspaces that they physically occupy or are necessary for their access or utilization. The location of these objects will determine the areas that are open for navigation and those that must be avoided.

Meanwhile, the meanings of O-Space, R-Space and F-Space have been established. O-Space is a non-navigable subspace that SM-objects physically inhabit. R-Space refers to the area of space that is open to navigation and can be traversed at will. In contrast, F-Space is a subspace functionally occupied by SM-objects or motion-less M-objects for user interactions. To be more exact, F-Spaces are spaces that are not considered navigable unless they form an essential part of the navigation system.

It is essential to have a solid understanding of how a structure would behave in the event of a fire. The establishment of minimum construction criteria is done to

assist in maintaining the building's structural integrity for the amount of time necessary for evacuating the building or moving to a secure position within the building. The rate at which a fire will spread is directly proportional to the combustibility of the material it is burning. These two components are equally crucial to ensuring that both lives and property are not lost in a fire.

A building's structural integrity can be affected by a fire. Different materials have varying degrees of resistance to fire [18]. When there is a risk to the route's stability, it is impossible to travel along that route. When the length of the fire is greater than or comes dangerously close to exceeding the fire resistance duration of the building material, the stability of the route is jeopardized. Unsecured shafts or openings pose an additional risk to the safety of the firefighters and the operation as a whole, and the planning of the route should take care to avoid them.

The amount of smoke in a region affects vision, which, in turn, decides whether or not a path may be used. If a path is devoid of smoke, has adequate visibility, and is easily navigable. In certain circumstances, smoke will only be found above a predetermined height beneath the ceiling. The area below the smoke may then be free of smoke and suitable for passage.

The NFPA (National Fire Protection Association) 220 Standard on Types of Building Construction [19] defines the different types of building construction by basing it on the combustibility and the fire resistance rating of the structural parts of a building. When we speak of materials or assemblies having a specific fire resistance rating, we refer to the amount of time measured in minutes or hours that they have been able to withstand exposure to fire, as established by specific tests. In this study, we take combustibility as a factor in determining the definition of a static object by taking the categorization from NFPA 220 [19].

#### A. Definition 3.1

(Redefinition of S-objects): static (S-objects) as objects that can neither move by themselves nor be moved (e.g., construction elements such as walls, columns, stairs, etc.) [10]. In this paper it can be redefined as three types of objects based on combustibility:

- Scm (Static-Combustible) is a material that, in its intended form and under the anticipated conditions, will ignite and burn;
- Slcom (Static-limited combustible) is a material does not meet the standards for non-combustible material.
- Sncom (Static-non-combustible), in its intended form and under the predicted conditions, the material will not ignite, burn, support combustion, or emit flammable gases when exposed to fire or heat.

Still in the FSS [10], the agents are the primary dynamic actors that are considered. During the process of estimating a route and when actually navigating, the A-Spaces intend to take into account the dimensions of the objects as well as the required amount of space around them. An 'A-Space' is a

clearance space containing one or more agents and the SM objects they carry if they have any.

This clearance space can be characterized as A-Space. However, the study [10] only considered the available free space. In rescue operations carried out by first responders, free space is taken into account, but the most important aspect is the space that is considered safe from the worst possibilities, such as fire and other hazards. The study [20] described a safety zone as "an area identified by qualities that give freedom from danger, harm, or injury". In definition 3.2, we are redefining A-Spaces by considering the conditions that may occur in rescue operations.

### B. Definition 3.2

(Redefinition of A-Spaces): we re-identify two types of A-Spaces:

- a safe one (AS-Space) corresponding to the remaining space that give freedom from danger;
- a non-safe one (ANS-Space) corresponding to the remaining space that not give freedom from danger.

We took a case by taking a 3D floor plan of the 2nd floor of the Westport House from WRLD3D.com, and by adding a heatmap, we simulated an unsafe area caused by a fire. As seen in Fig. 2, the area given the red heatmap is ANS-Space, while the rest is AS-Space.



Fig. 2. AS-Space and ANS-Space within indoor building.

## IV. SPACE SUBDIVISION AND NETWORK DERIVATION

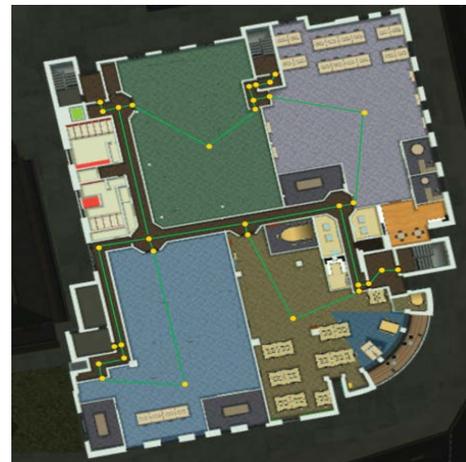
The navigation network is a crucial component of most navigation and evacuation approaches[21]. A navigation network in indoor navigation refers to a system of interconnected paths or routes that can be used to navigate through an indoor environment, such as a building. The navigation network may include information on the layout of the building, the location of exits, stairwells, and elevators, and the presence of hazards or obstacles. It may also include information on the characteristics of the spaces within the

building, such as the combustibility of the materials, the presence of windows, or the availability of emergency exits.

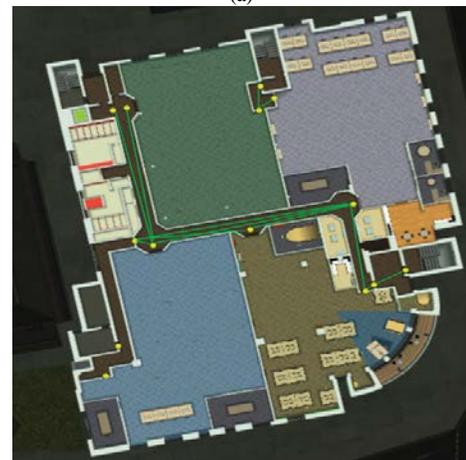
There are widely accepted methods for developing indoor navigation networks, such as medial axis transformation (MAT), visibility graph (VG), or mixtures of them [22]. In other studies, CDT (Constrained Delaunay Triangulation)[23], and generalized Voronoi diagram (GVD) are also used. In this study we use MAT and VG to create a navigation network.

Straight skeleton is the process of skeletonizing a geometric space. The straight MAT algorithm [24] generates a 3D indoor skeletal graph. Fewer vertices are added to the network, resulting in increased computing efficiency. Taneja et al. in [25] employed direct MAT to turn IFC-based data into a geometric topology network, which might serve as a navigational aid model. As can be seen in Fig. 3(a), we created a network using MAT based on the existing floor plan. While in Fig. 3(b) is the network derivation using VG.

Deriving nodes from existing building plans is the first and most essential step in developing an VG (visibility Graph)[26]. Based on their navigational functions, there are primarily two types of nodes in a VG [27]. First, a junction node denotes the intersection of at least two corridor segments. The second, portal nodes describe wall openings, such as those occupied by doors and windows in each room.



(a)



(b)

Fig. 3. Navigation network: a) MAT network and (b) VG network.

## V. PATH PLANNING

Safe routes are essential in indoor navigation for first responders because they help to ensure that the responders can reach their destination safely and efficiently, even in hazardous or challenging environments. This is especially important in emergency situations, where time is of the essence and responders may need to navigate through unfamiliar or potentially dangerous spaces. Common path findings utilizing typical navigation methods may not be sufficient in an emergency, as they may be too dangerous or unavailable due to damage. Critical in these situations is the ability of emergency and rescue professionals to find other and best routes [28].

In this study, we propose an algorithm for determining the safest path for first responders by taking into account factors such as the combustibility of materials and dynamic changes in the building such as fire. To create an algorithm for safe indoor navigation for multiple first responders that takes into account the combustibility of materials, dynamic changes in the building such as fire, and is based on space subdivision, we could consider the following steps:

- 1) Identify the start and end points of the route for each first responder.
- 2) Obtain a map of the building or structure, including information on the layout and location of rooms and corridors.
- 3) Divide the building or structure into smaller spaces, such as rooms and corridors.
- 4) Determine the combustibility of materials in each space. This information can be obtained through building codes or by consulting with the building's owner or management.
- 5) Use the map and combustibility information to identify potential hazards in each space, such as rooms or corridors with highly combustible materials.
- 6) Consider the potential for structural collapse and other hazards that could block the route such as smoke.
- 7) Based on this information, determine safe routes for each first responder that avoids or minimizes exposure to potential hazards. This can be done by selecting paths through the building or structure that avoid spaces with high hazards and maximize the use of spaces with low hazards.

Here is the pseudocode of our proposed algorithm:

```
PROCEDURE findSafeRoutes(starts, ends, map, combustibility)
    spaces <- list of all spaces in map
    routes <- empty list
    FOR EACH start, end IN starts, ends
        route <- empty list
        currentSpace <- start
        WHILE currentSpace != end
            nextSpace <- null
            lowestRisk <- infinity
            FOR EACH neighbor IN neighbors of currentSpace
                totalDistance <- distance between currentSpace and
```

```
neighbor
                totalHazardExposure <- 0
                riskOfCollapse <- 0
                IF neighbor has highly combustible materials
                    totalHazardExposure <- totalHazardExposure + 1
                IF neighbor is at risk of structural collapse
                    riskOfCollapse <- riskOfCollapse + 1
                IF neighbor is on fire or has high levels of smoke
                    totalHazardExposure <- totalHazardExposure + 10
                IF neighbor is a window and is suitable for exiting the
building
                    totalHazardExposure <- totalHazardExposure - 5
                totalRisk <- totalDistance + totalHazardExposure +
riskOfCollapse
                IF totalRisk < lowestRisk
                    lowestRisk <- totalRisk
                    nextSpace <- neighbor
            route <- route + (currentSpace, nextSpace)
            currentSpace <- nextSpace
        routes <- routes + route
    RETURN routes
```

This pseudocode defines a procedure `findSafeRoutes` that takes as input the start and end points of the routes for each first responder, a map of the building or structure, and information on the combustibility of materials in the building. The procedure returns a list of routes as output. The algorithm determines the safest route for each first responder by traversing the spaces in the building or structure, calculating the total distance, total hazard exposure, and risk of structural collapse for each possible next step, and selecting the step with the lowest total risk as the safest route. If a space is on fire or has high levels of smoke, the algorithm increases the total hazard exposure to prioritize avoiding these spaces. If a space is a window that is suitable for exiting the building, the algorithm decreases the total hazard exposure to prioritize using this space as an alternative exit. The algorithm continues this process until the end point is reached.

## VI. EXPERIMENT

An agent-based simulation is a well-established technique for modeling various applications, including evacuations[29]. ABM (Agent-based modelling) is a simulation technique in which an entity functions as an agent. The rules set, interactions with other agents and the surrounding environment determine the behavior of individual agents. These simulations can be used to test and evaluate different evacuation strategies, communication protocols, and other emergency response procedures in a safe and controlled environment. The simulation can be used to test different scenarios and evaluate the performance of the first responders. It can also help identify bottlenecks in the evacuation process and evaluate the effectiveness of different communication and coordination strategies. The results of the simulation can be used to improve training for first responders and to develop more effective emergency response plans.

In this research, Pathfinder is combined with Pyrosim. The smoke display file is imported into the FDS simulation

parameter data. FDS integration and passenger delay in smoke were investigated to allow ABM to limit agent speed due to smoke. In this study, the experiment was carried out in a six-story building, as seen in the Fig. 4.

The data we use uses the floor plan provided on Pathfinder. First, the IFC file is imported into Pyrosim to then create a building geometry model. In Pyrosim, we set a fire point with a power of 800.0 kW/m<sup>2</sup>, which is placed in a room on the 3rd floor of the building as shown in Fig. 4.

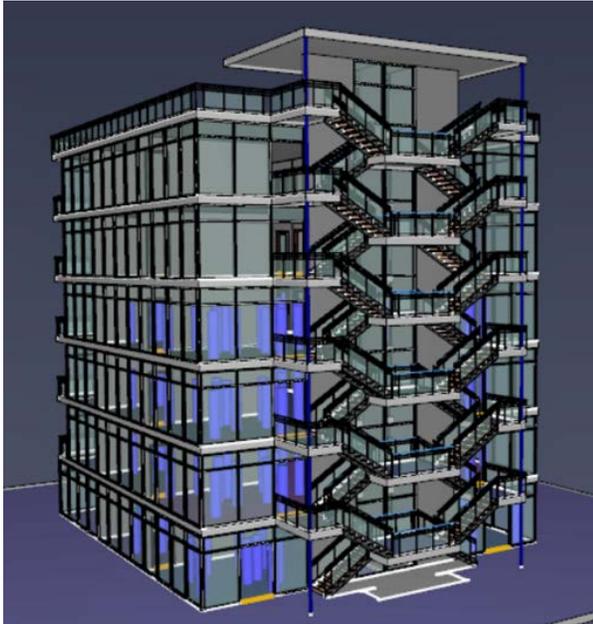


Fig. 4. Case study buildings.

Four visibility devices were installed in the building to record changes in the building's smoke levels. The bigger the smoke, the higher the total hazard exposure, as in the algorithm we proposed, which will increase the total risk.

In Fig. 5 we can see a fire simulation that occurred in the building. Smoke can affect the walking speed of first responders during an indoor evacuation by limiting visibility and making it difficult to navigate through the building. Smoke inhalation can also cause respiratory distress, which can slow down first responders. Additionally, the heat generated by a fire can make it physically challenging for first responders to move quickly through the building. This can be dangerous as it can cause them to become disoriented and lost, making it more difficult to evacuate the building and rescue those inside.

Visibility will be used to slow down people's walking speed. To achieve this, visibility will be measured in several areas within the model. The measurement location will be along the evacuation route, preferably stored on the floor where the fire occurred. Using the visibility distance measured at the location, the local velocity factor will be generated as a function of time. Responders will then reduce their speed and modify their route accordingly. [30] provides a walking speed function as a function of visibility. This is referred to as absolute walking speed in the paper, but we will treat it as a factor that slows the speed of each occupant. Fig. 6 depicts a

linear relationship between walking speed and visibility, which can be expressed by the Eq. (1):

$$\text{Walkingspeed} = \min(1, \max(0.2, 1 - 0.34 * (3 - \text{vis}))) \quad (1)$$

In Fig. 7, we can see the extracted navigation mesh from one of the floors of the building. This navigation mesh will be useful for responders in determining routes.

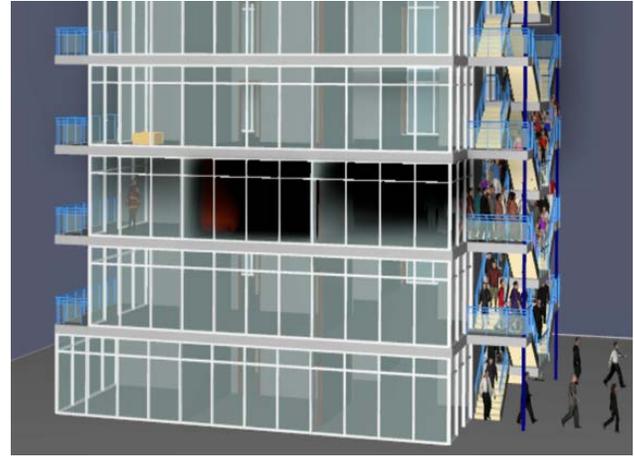


Fig. 5. Fire simulation scene in the building.

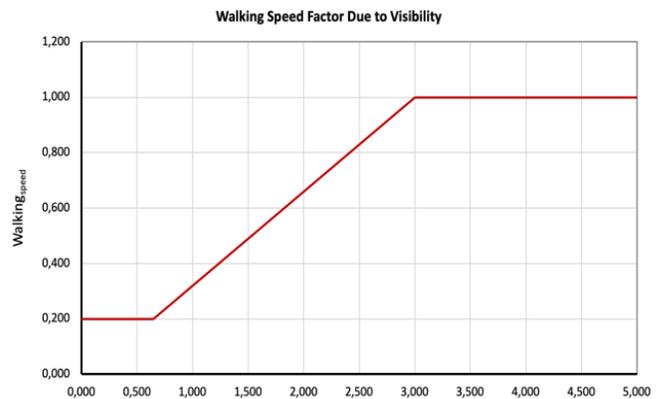


Fig. 6. Walking speed factor due visibility.

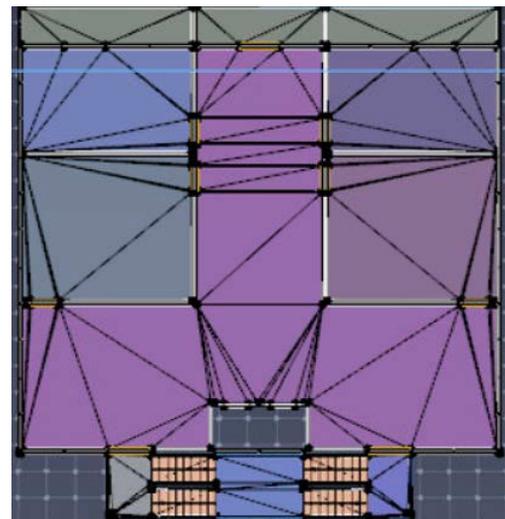


Fig. 7. Extracted navigation mesh.

Navigation mesh space subdivision is a method of breaking down an indoor environment into smaller, manageable sections for the purpose of pathfinding and navigation. This can be particularly useful for first responders, who may need to navigate through large, complex indoor environments such as buildings, airports, or shopping malls. By breaking down the environment into smaller sections, navigation mesh space subdivision can help first responders to navigate through an indoor environment more efficiently and safely by reducing the complexity of the environment and allowing them to focus on specific areas.

The simulation of route choice in this model employs a locally quickest path planning approach. This approach involves ranking routes hierarchically based on local information such as the location of people and the waiting times at exits.

Our model uses two profiles, Occupants and Responders. The number of occupants of the building is 360 people, and 90 people inhabit each floor. Occupants were only set for four floors, while for the floor where the fire occurred, five first responders were simulated during the experiment. The decision to set occupants for only four floors and to simulate five first responders was made to maintain a manageable size for the simulation, while still allowing for a realistic evaluation of the evacuation process and the response of first responders to the emergency situation. By limiting the number of floors with occupants, the experiment could focus on the behavior of the occupants in response to the emergency and the effectiveness of the evacuation procedures. The simulation of five first responders on the floor where the fire occurred was intended to evaluate their ability to navigate the space, locate the occupants, and carry out the evacuation process. Overall, the decision to set occupants for only four floors and simulate five first responders on the floor where the fire occurred was a deliberate choice to ensure a manageable size for the experiment while still providing a realistic evaluation of the evacuation and emergency response processes.

The occupants are assigned a profile with a normal speed of 0.8 m/s to 1.2 m/s. These profiles have a zero Priority Level. The responders' profile has a constant velocity of 1.19 meters per second and a Priority Level of 1. This means in the event of a conflict during movement, and responders will be given priority. However, when going through a smoke-filled room, the pace will decrease according to the value acquired by the visibility tool and then calculated based on the walking speed factor.

### VII. RESULTS AND DISCUSSION

This section summarizes the findings of the fire simulation, with a focus on the first responders, based on the assumptions provided in the previous section. Visibility measurement tools output data to csv files. An example of measurement result data can be seen in Fig. 8. The data was generated from four visibility measurement devices during the simulation run. After the simulation, the data from the visibility measurement devices is processed using the equation shown in the Speed walking Factor to calculate the velocity factor as a function of time based on visibility.

1	s	m	m	m	m
2	Time	visibility1	visibility2	visibility3	visibility4
214	42.20	1.50	18.89	15.50	30.00
215	42.47	1.49	19.16	15.19	30.00
216	42.61	1.48	19.43	14.89	30.00
217	42.86	1.48	19.69	14.59	30.00
218	43.04	1.47	19.94	14.23	30.00
219	43.24	1.47	20.11	13.90	29.88
220	43.43	1.47	20.24	13.53	28.54
221	43.63	1.47	20.34	13.16	27.13
222	43.82	1.47	20.41	12.81	26.00

Fig. 8. Result data from visibility device measurements.

Fig. 9 shows a graph of the velocity factor resulting from the measurements.

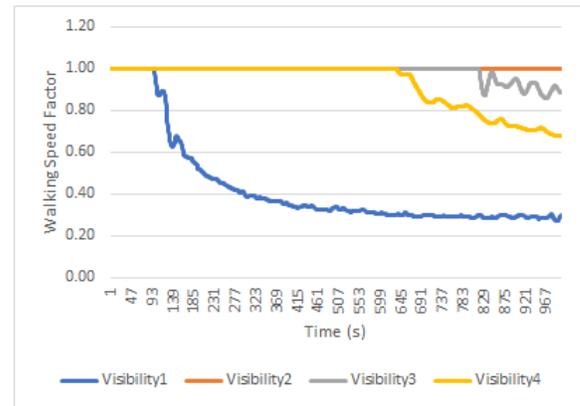


Fig. 9. Walking speed factor.

The time taken by the five responders to carry out the evacuation operation to the location where the fire/smoke occurred was measured during the simulation. The time taken for the responders to enter the building and return safely was calculated, and the results are presented in Table I. Table I compares the time required using the walking speed factor to the time that does not take the walking speed factor into account.

Table I and Fig. 10 shows that the smoke will affect the walking speed of each responder, where the speed will be slower according to the amount of smoke calculated from the visibility devices. In Table I, a non-factor column is an event where a fire does not occur and does not occur immediately. So that the walking speed of first responders is faster than when there is a fire. Smoke can greatly impact the visibility and the walking speed of first responders during indoor emergency operations. Smoke can obscure visibility, making it difficult for first responders to navigate through the environment and locate victims or exits. It also can cause respiratory issues and make it harder for first responders to move quickly through the environment. The smoke can greatly affect the walking speed of first responders, making it more difficult for them to navigate through the environment and locate victims or exits. This can have serious consequences in emergency situations, as it can delay response times and make it more difficult to rescue victims or contain fires.

TABLE I. COMPARISON OF EXIT TIME

Name	Exit Time (s)	
	Non-Factor	With Factor
Responder1	234.825	218.35
Responder2	188.275	148.275
Responder3	239.05	239.625
Responder4	226.025	219.6
Responder5	244.075	227.075

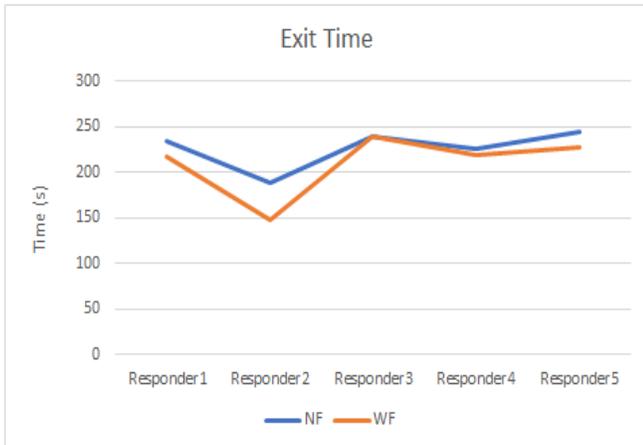


Fig. 10. Exit time from responders.

### VIII. CONCLUSION AND FUTURE WORK

The present study introduces a new framework for space subdivision in indoor navigation specifically developed for first responders. The framework expands upon previous research by including the combustibility factor as a key variable to consider in the event of a fire disaster in a building. Additionally, definitions of safe and unsafe areas for first responders were incorporated based on factors such as the level of hazard and smoke concentration. From these definitions, an algorithm was developed to accommodate the framework, which was then evaluated using Pyrosim and Pathfinder software. Pyrosim was utilized to create a fire model that was measured using a visibility device. The resulting measurements were then subjected to a walking speed factor calculation, which affected the path and walking speed of the first responders. The proposed framework enhances the response time and increases the chances of successful evacuation for first responders, ultimately improving their safety and effectiveness during rescue operations in indoor environments.

Unlike previous studies on indoor navigation, the proposed framework captures dynamic changes such as smoke levels that can impact the navigation path and walking speed of first responders. The experimental results demonstrate that the five simulated first responders were able to complete the evacuation process within 148 to 239 seconds when the fire occurred at a specific location in the building. By accounting for dynamic changes in the indoor context, the proposed framework can identify suitable navigation paths and safe areas for first responders that may not have been considered in

previous research. Thus, the proposed framework represents a significant improvement over previous studies, and has the potential to enhance the safety and effectiveness of first responders in emergency situations.

Future works from this research: integrating the proposed framework with other technologies, such as augmented reality, virtual reality, and internet of things (IoT), to enhance the navigation experience for first responders. Enhancing the cost function: The study should consider more factors in the cost function such as the time consumed by the first responders, the physical strain on them, and the risk of injury. Real-time monitoring: The study could explore the implementation of real-time monitoring systems to track the first responders' movements and provide them with real-time guidance.

### REFERENCES

- [1] N. E. Klepeis et al., "The National Human Activity Pattern Survey (NHAPS): a resource for assessing exposure to environmental pollutants," no. September 1998, 2001.
- [2] A. A. Diakit , S. Zlatanova, and K. J. Li, "ABOUT the SUBDIVISION of INDOOR SPACES in INDOORGML," *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 4, no. 4W5, pp. 41–48, 2017, doi: 10.5194/isprs-annals-IV-4-W5-41-2017.
- [3] U. S. F. Administration and N. Fire, "Fire-Related Firefighter Injuries Reported to the National Fire Incident Reporting System ( 2015-2017 )," 2019.
- [4] G. Prati and L. Pietrantonio, "The relation of perceived and received social support to mental health among first responders: A meta-analytic review," *J Community Psychol*, vol. 38, no. 3, 2010, doi: 10.1002/jcop.20371.
- [5] M. P. Kwan and J. Lee, "Emergency response after 9/11: The potential of real-time 3D GIS for quick emergency response in micro-spatial environments," *Comput Environ Urban Syst*, vol. 29, no. 2, pp. 93–113, 2005, doi: 10.1016/j.compenvurbsys.2003.08.002.
- [6] Y. Zheng, M. Peter, R. Zhong, S. O. Elberink, and Q. Zhou, "Space subdivision in indoor mobile laser scanning point clouds based on scanline analysis," *Sensors (Switzerland)*, vol. 18, no. 6, pp. 1–20, 2018, doi: 10.3390/s18061838.
- [7] A. I. HADIANA, S. S. K. BAHARIN, and N. S. HERMAN, "Space subdivision for indoor navigation: A systematic literature review," *J Theor Appl Inf Technol*, vol. 98, no. 15, pp. 3093–3105, 2020.
- [8] W. Zhen, Z. Zuo, M. P. Kwan, L. Yang, S. Zhou, and H. Qian, "Capturing dynamic navigable space: an interactive semantic model to expand functional space for 3D indoor navigation," *International Journal of Geographical Information Science*, vol. 0, no. 0, pp. 1–25, 2022, doi: 10.1080/13658816.2022.2048387.
- [9] H. Wu, H. Yue, Z. Xu, H. Yang, C. Liu, and L. Chen, "Automatic structural mapping and semantic optimization from indoor point clouds," *Autom Constr*, vol. 124, no. October 2020, 2021, doi: 10.1016/j.autcon.2020.103460.
- [10] A. A. Diakit  and S. Zlatanova, "Spatial subdivision of complex indoor environments for 3D indoor navigation," *International Journal of Geographical Information Science*, vol. 32, no. 2, pp. 213–235, 2018, doi: 10.1080/13658816.2017.1376066.
- [11] U. Isikdag, S. Zlatanova, and J. Underwood, "A BIM-Oriented Model for supporting indoor navigation requirements," *Comput Environ Urban Syst*, vol. 41, pp. 112–123, 2013, doi: 10.1016/j.compenvurbsys.2013.05.001.
- [12] J. Shang, X. Tang, and F. Yu, "A Semantics-based Approach of Space Subdivision for Indoor Fine-grained Navigation," *Journal of Computational Information Systems*, vol. 11, no. April 2016, 2015, doi: 10.12733/jcis14367.
- [13] S. Zlatanova, L. Liu, G. Sithole, J. Zhao, and F. Mortari, "Space subdivision for indoor applications," 2014.

- [14] M. Krūminaitė and S. Zlatanova, "Indoor Space Subdivision for Indoor Navigation," in Proceedings of the ACM International Symposium on Advances in Geographic Information Systems, 2014, pp. 25–31.
- [15] W. Zhen, L. Yang, M. P. Kwan, Z. Zuo, H. Qian, and S. Zhou, "Generating comfortable navigable space for 3D indoor navigation considering users' dimensions," *Sensors (Switzerland)*, vol. 20, no. 17, pp. 1–25, 2020, doi: 10.3390/s20174964.
- [16] A. A. Diakitė and S. Zlatanova, "Spatial subdivision of complex indoor environments for 3D indoor navigation," *International Journal of Geographical Information Science*, vol. 32, no. 2, pp. 213–235, 2018, doi: 10.1080/13658816.2017.1376066.
- [17] M. Aleksandrov, D. J. Heslop, and S. Zlatanova, "3D Indoor Environment Abstraction for Crowd Simulations in Complex Buildings," *Buildings*, 2021.
- [18] A. Dugstad, S. Bralić, and J. Abualdenien, "Path planning through disaster scenes : qualitative interviews to assess relevant parameters," in Proc. of the 32th Forum Bauinformatik, 2021.
- [19] National Fire Protection Association, "NFA 220 Standar on Types of Building Construction," 2018.
- [20] M. Beighley, "Beyond the safety zone: creating a margin of safety," *Fire Management Today*, vol. 55, no. 4, 1995.
- [21] P. Boguslawski, S. Zlatanova, D. Gotlib, M. Wyszomirski, M. Gnat, and P. Grzempowski, "3D building interior modelling for navigation in emergency response applications," *International Journal of Applied Earth Observation and Geoinformation*, vol. 114, no. October, 2022, doi: 10.1016/j.jag.2022.103066.
- [22] M. Fu, R. Liu, B. Qi, and R. R. Issa, "Generating straight skeleton-based navigation networks with Industry Foundation Classes for indoor way-finding," *Autom Constr*, vol. 112, no. December 2019, p. 103057, 2020, doi: 10.1016/j.autcon.2019.103057.
- [23] F. Mortari, E. Clementini, S. Zlatanova, and L. Liu, "An indoor navigation model and its network extraction," *Applied Geomatics*, vol. 11, no. 4, pp. 413–427, 2019, doi: 10.1007/s12518-019-00273-8.
- [24] J. Lee, "A spatial access-oriented implementation of a 3-D GIS topological data model for urban entities," *Geoinformatica*, vol. 8, no. 3, pp. 237–264, 2004, doi: 10.1023/B:GEIN.0000034820.93914.d0.
- [25] S. Taneja, B. Akinci, J. H. Garrett, and L. Soibelman, "Algorithms for automated generation of navigation models from building information models to support indoor map-matching," *Autom Constr*, vol. 61, pp. 24–41, 2016, doi: 10.1016/j.autcon.2015.09.010.
- [26] Z. Zhou, R. Weibel, K. F. Richter, and H. Huang, "HiVG: A hierarchical indoor visibility-based graph for navigation guidance in multi-storey buildings," *Comput Environ Urban Syst*, vol. 93, p. 101751, 2022, doi: 10.1016/j.compenvurbsys.2021.101751.
- [27] L. Yang and M. Worboys, "Generation of navigation graphs for indoor space," *International Journal of Geographical Information Science*, vol. 00, no. 00, pp. 1–20, 2015, doi: 10.1080/13658816.2015.1041141.
- [28] P. Boguslawski, L. Mahdjoubi, V. Zverovich, and F. Fadli, "Automated construction of variable density navigable networks in a 3D indoor environment for emergency response," *Autom Constr*, vol. 72, pp. 115–128, 2016, doi: 10.1016/j.autcon.2016.08.041.
- [29] S. T. S. Abadi, N. M. Tokmehdash, A. Hosny, and M. Nik - bakht, "Bim - based co - simulation of fire and occupants' behavior for safe construction rehabilitation planning," *Fire*, vol. 4, no. 4, 2021, doi: 10.3390/fire4040067.
- [30] K. Fridolf, D. Nilsson, H. Frantzich, E. Ronchi, and S. Arias, "WALKING SPEED IN SMOKE: REPRESENTATION IN LIFE SAFETY VERIFICATIONS," in Conference: The 12th International Conference on Performance Based Codes and Fire Safety Design MethodsAt: Oahu, Hawaii, 2020, vol. 5, no. 3, pp. 248–253.

# Leaf Diseases Identification and Classification of Self-Collected Dataset on Groundnut Crop using Progressive Convolutional Neural Network (PGCNN)

Anna Anbumozhi, Shanthini A

Department of Data Science and Business Systems  
SRM Institute of Science and Technology, Kattankulathur, Chennai 603203

**Abstract**—A healthy crop is required to provide high-quality food for daily consumption. Crop leaf diseases have more influence on agronomic production and our country. Earlier, many scholars relied on traditional techniques to detect and classify leaf diseases. Furthermore, classification at an early stage is impossible when there are not enough experts and inadequate research facilities. As technology progresses into our day to day life, an Artificial Intelligence subset called Deep Learning (DL) models plays a vital role in the automatic identification of groundnut leaf diseases. The essential for controlling diseases that are spread to the healthy development of groundnut farming. Deep Learning can resolve the issues of finding leaf diseases early and effectively. Most of the researchers concentrate on detecting leaf diseases by doing research in Machine Learning (ML) approaches, which leads to low accuracy and high loss. To achieve better accuracy and decreases the loss in the DL model by identifying the leaf diseases of groundnut crops at an early stage, we propose the Progressive Groundnut Convolutional Neural Network (PGCNN) model. This paper mainly focuses on identifying and classifying groundnut leaf diseases with a self-collected dataset which is collected from the various climatic conditions around the village located nearby Pudukkottai district, Tamil Nadu, India. The common diseases that occurred in those areas were gathered namely early spot, late spot, rust, and rosette. Model Performance metrics analysis was done to evaluate the model performance and also compared with the various DL architectures like AlexNet, VGG11, VGG13, VGG16, and VGG19. The proposed models have achieved a training accuracy of 99.39% and a validation accuracy of 97.58%, continuing with an overall accuracy of 97.58%.

**Keywords**—*Leaf Diseases Identification (LDI); Progressive Groundnut Convolutional Neural Networks (PGCNN); Self-Collected Dataset; AlexNet; VGG Models*

## I. INTRODUCTION

Agriculture is a major source of revenue for the farmers in India. Despite the fact that agriculture no longer makes up the majority of the country's Gross Domestic Product (GDP) and that other industries' contributions have grown more quickly, agricultural productivity has increased. Timely identification of crop leaf diseases is critical in agricultural yield, quality management, and decision-making. [1] This research builds the model based on CNN architecture for groundnut leaf diseases with five categories which was common in Gujarat area with the progressive resizing using cross entropy loss

function and achieved the results as 96.12% accuracy. The peanut, scientifically known as *arachis hypogaea*, it is a herbaceous yearly plant that belongs to the Fabaceae family and is cultivated for its oil, edible nuts, and nutritious snacks. The symptoms that are shown in the leaf diseases can be easily differentiated between the two infections based on their appearance, spot color, and shape. However, farmers must suffer as a result of inadequate agricultural income. Plants that produce peanuts are typically quite small and have slender stems and leaves that resemble feathers. The leaves are attached to the stalk in a manner that resembles a leaf and is arranged in pairs that alternate. [2] Describes the end-to-end Internet of Things (IoT) constructed system for groundnut leaf disease detection and castor oil plant leaf disease recognition. The vast majority of peanuts grown for commercial purposes are crushed up to extract their oil, which is then put to use in culinary endeavours. A pressed cake is produced as a by-product of oil extraction, and in addition to its use as animal feed, it is also put to use in the manufacturing of peanut flour. Raw kernels are frequently roasted into a snack food and consumed by roasted or boiled manner.

Here, the four types of diseased leaves were selected to experiment with disease identification: early spot, late spot, rust, and rosette. These four have been taken as unhealthy class for model creation in this paper. The above leaf disease images were gathered from the area namely Pudukkottai, Tamil Nadu, India in various climates. All the images are captured with a smartphone-branded VIVO V2029 model with a camera quality of 13 Mpx LED. Gathered images are transferred to the system for further processing, whether the leaf images were affected by diseases or not. To identify groundnut leaf diseases automatically we are proposing Progressive Groundnut Convolutional Neural Networks (PGCNN). The study [3] represents the idea for the recognition and classification on groundnut leaf diseases using the backpropagation procedure by the color transformation in leaves. Previously the traditional CNN architecture was used in many different research papers to identify leaf diseases but there no accurate identification was found. So, to address that problem we are developing the model with a progressive convolutional stack with customized pooling layers to reduce the size of the image and activation function as sigmoid for binary classification.

Convolutional neural networks (CNNs) have been demonstrated to perform well in a variety of image

classification tasks. We apply CNNs to the problem of identifying and classifying peanut leaf diseases in this paper. Peanut is a key food crop in many regions of the world, and leaf diseases can drastically limit productivity. Early and accurate disease diagnosis is critical for effective disease control. However, due to a large number of diseases and the similarity of symptoms among them, diagnosing peanut leaf diseases is frequently difficult. CNNs have been demonstrated to be effective in a variety of image classification tasks. Our CNN model achieves 96.7% accuracy on a held-out test set, which is much greater than the accuracy of a baseline model that uses traditional hand-crafted features.

The rest of the paper is sectioned by the following pattern: Section II details about the existing methods that are used for identifying and classifying the leaf diseases as literature review with various peer review journals; Section III describes about the datasets which was collected from the agricultural field from Pudukkottai district and also shows the sample leaves of healthy and unhealthy leaves; Section IV discusses the details of flow diagram of dataset and the proposed PGCNN model for leaf disease identification and classification among the groundnut self-collected dataset; Section V explains about the implementation details of PGCNN model and the results were achieved, as well as the discussion that was pertinent to them are presented; Section VI shows the results and discussion about the PGCNN model and plots the graphs for training and validation accuracy; Section VII compares our proposed model to the existing CNN architecture to show the proposed model works well than the existing model; and Section VIII concludes the paper with briefs about the proposed model results with future enhancement.

## II. LITERATURE REVIEW

### A. Identification of Diseases

In recent days, the identification and Classification of groundnut crop leaf diseases are done automatically by ML and DL algorithms. Identification of healthy or diseased leaves on groundnut crops is based on the features extracted from the leaves as yellow or brown changes in color from the original green texture of normal groundnut leaves. After identifying the healthy and diseased leaves, it has to be classified whether it is infected or not. Some of the research papers were taken as reference papers for this work to attain the solution with a Deep Learning model for the evaluation of leaf diseases on groundnut crops.

In [4] researcher has done the experiment to identify the Maize leaf diseases based on the feature enhancement and designed a neural network with the base as Alexnet architecture. DMS-Robust Alexnet architecture was proposed with dilated and multi-scale convolution to improve the quality of feature extraction. To avoid overfitting batch normalization is used, to improve the convergence and accuracy researcher used the activation function as PRelu, and Optimizer as Adabound and achieved 98.62% with DMS-Robust Alexnet.

In [5] author proposed a model with a deep learning approach called the Inception model and Rainbow

concatenation (INAR-SSD) model which is based on Improved Convolutional Neural Networks for apple disease detection. The major five types of leaf diseases were selected and implemented in the INAR-SSD model with a recognition performance of 78.80% in Apple Leaf Disease Dataset (ALDD).

The research [6] proposed the Leaf-GAN model to create four different types of leaf diseases for grape crops to identify the healthy or diseased images and fed the generated images to training. There were total of 4062 images before using Leaf GAN, after using augmentation techniques 8124 images are generated to reduce the overfitting problem. Upon CNN models Xception attains an accuracy of 98.70% on testing the datasets.

The study [7] proposed a fine-grained GAN method for identification of grape leaf spot disease to improve the training images to get good performance accuracy and also faster R-CNN is integrated with the above method. With the proposed method higher accuracy was achieved with Resnet-50 as 96.27%.

The study [8] achieved 98.75% and 96.25% of accuracy by implementing the pretrained CNN model called AlexNet and GoogleNet models. This model performed better than the traditional pattern recognition techniques. Here the researcher used five-fold cross-validation approach and crop selected for the plant disease identification is soybean.

The author [9] proposed Restructured Residual Dense Network (RRDN) for identification of tomato leaf diseases with datasets of AI challenger 2018. With residual and dense networks, which minimize the amount of parameters to increase computation accuracy, the result was 95% with top-1 average accuracy.

In [10] the study achieved top-1 level accuracy 94.33% by experimenting Deep Convolutional Generative Adversarial Network (DCGAN) to improve the tomato leaf disease recognition that increases the generalization ability. This network improves the performance, decreases the dataset collection cost and also establishes the diversity of generalization of DCGAN recognition models.

Also, [11] Proposed Deep Convolutional Neural Networks with multiclass classifier for detecting common rice crop diseases with the base model of AlexNet by SGD optimizer on the learning rate of 0.0001 and achieved 91.23% accuracy.

### B. Classification of Leaf Diseases

The research [12] proposed the pretrained model of modified InceptionResNet-V2 (MIR-V2) based on CNN by transfer learning to identify the illness of tomato leaves. This model is trained with both the public and custom dataset of 7 types of diseases. Model achieved the accuracy of 98.92% and F1-Score is 97.94%.

The research work in [13] proposed new model which is based on recognition and classification of groundnut crop leaf diseases. ICS algorithm is used for segmenting the leaves which are affected by diseases. Then MSO algorithm has been used for multiclass feature extraction and MO-DNN algorithm is used for disease classification of multi-classes. At last,

GLD-HML was proposed for analyzing the benchmark datasets to show the performance metrics achieved better than existing.

The authors in [14] used two pre-trained models namely EfficientNetB0 and DenseNet121 for feature extraction from the corn leaves to identify and classify the diseases. The results achieved by new model is 98.56% and compared with two existing model called ResNet152 and InceptionV3 achieved accuracy of achieved accuracy of 7% and 96.26% respectively.

In [15] researcher proposed the Deep Convolutional Neural Network (DCNN) which is based on the evolved concept of transfer learning. Here three types of optimizers were verified with the new model for tomato leaf diseases to identify which optimizer (Adam, SGD and RMSprop) works well for the model. The experimental results showed that the transfer learning model attained the good accuracy with Adam optimizers.

The study [16] utilized the DCNN model for determination and classification of groundnut crop leaves with different types of diseased leaves. This paper achieves accuracy of 95.28% with the optimizer of stochastics gradient with momentum method in DCNN model. Overall accuracy for the proposed model delivers 99.88%.

The research [17] proposed the new model for classifying the groundnut crops leaf as healthy and unhealthy with CNN algorithm with the image compression techniques as DCT, DFT and DWT. The results show that the overall computational time was reduced by comparing the CNN with ResNet50 architecture.

The author [18] created the method for precise detection and classification of groundnut leaf diseases, the method which is proposed by the author, algorithms named H2K which combines the strengths of the three-concept called Harris corner detector, the HOG, and the K-Nearest Neighbor classifier.

In [19] the researcher developed the system for the detection and classification of nutrient deficiency in groundnut leaf especially on nitrogen level. DCNN was implemented for the dataset to achieve the high accuracy and also achieved 95% for training, 92% for validation result.

### III. GROUNDNUT DATASETS

In the section we are going to describe about the groundnut dataset where we have collected from Pudukkottai district, Tamil Nadu, India. Here the detailed description of groundnut leaves and total number of leaves collected with various diseases occurred in the leaf with healthy leaf as on category. [20] explains about the disease-free plant growth which produces more productivity with ELM algorithm with normalization using the benchmark dataset and obtains the optimal learning and better generalization. The dataset is collected with the five types of diseases including healthy leaves to identify whether the leaves are affected by disease or not. In this paper, the commonly affected diseases are selected for the identification of the leaf diseases process. Because these are the diseases that were affected in the area where the

dataset was collected. Details and stages of leaf disease that occurred in groundnut crops are briefed as follows;

#### A. Dataset Description

First Early Leaf spot is a common disease that reduces yield significantly and can be found anywhere the groundnut is grown. This encompasses countries and regions such as the United States of America, Australia, Fiji, the Solomon Islands, Tonga and in Indian states such as Gujarat, Andra Pradesh, Karnataka, Tamil Nadu, and Maharashtra. It is estimated that the leaf spot causes a reduction in yields of at least fifty percent in countries located in the Pacific. Approximately one month after sowing, the first infection of early leaf spot leaf disease will take place. The symptoms manifest themselves on the leaf as a reddish-brown large spot that is not perfectly circular and is surrounded by a yellow halo. The lesion on the lower surface is a light brown color.

Secondly, we took Late Leaf Spot types for the identification of diseases. This type will have infections at the beginning approximately seven weeks after sowing and appear as mostly circular dark brown small spots without a yellow halo. Carbon black color will display in lower surface lesions caused by late leaf spots.

Thirdly, Plants that have been infected with rosette will have the symptoms of a solid clump or gnome shoots, and each will have a tuft of small leaves arranged in a rosette pattern. Chlorosis and mosaic mottling are characteristics of the plant. The infected plants will continue to be underdeveloped and produce the flowers, with the immature flower only a small percentage of the groundnut pegs will mature into nuts, and there will be no seeds produced.

Last, Rust disease infects every part of the plant that is above ground. In most cases, the diseases were discovered at the time of plants at six weeks old approximately. On the backside of the groundnut leaves, tiny eruptions that range in color from brown to dusty chestnut are known as uredosori.

TABLE I. TYPES OF GROUNDNUT DISEASES

Sl. No.	Disease Selected	Symptoms	Duration of disease attacks
1.	Early Leaf Spot	The reddish-brown spot upon the leaf is surrounded by yellow halo nature.	1 month from sowing the seed.
2.	Late Leaf Spot	Carbon black color that will occur under the surface of the leas.	7 weeks after sowing the seed.
3.	Rosette	Yellow color leaf with mottling of the foliage	2 or 3 weeks from sowing.
4.	Rust	Tiny pustules that range in color from brown to dusty	After completing 6 weeks
5.	Healthy	doesn't change its natural green color.	Same as the nature of groundnut leaf color as green

The Table I shows about the groundnut leves diseases name with symptoms and describes about the time when the diseases will attack the leaf.

### B. Early Leaf Spot (ELS)

- Fig. 1. shows the sample of early leaf spot diseases in groundnut crops are fungal diseases caused by the fungus *Cercospora arachidicola*. This disease can cause extensive leaf spots and defoliation in groundnut crops. About one month after sowing, this disease begins to affect the crop.



Fig. 1. Early leaf spot

- The disease affects all above-ground plant parts, especially leaves. The two infections' leaf symptoms differ in appearance, spot color, and shape. Both fungi damage the petiole, stem, and pegs. As infection spreads, lesions from both species merge and spotted leaves drop early. Severe infections impair nut quality and production.

### C. Late Leaf Spot (LLS)

- Fig. 2. represents the LLS fungal disease that affects groundnut crops. This disease is triggered by the fungus *Phaeoisariopsis personata* and is characterized by the thick brown or black dots on the leaves. The spots can vary in size and shape, and they may be surrounded by a yellow halo. The disease can cause the leaves to drop off the plant, which can reduce the plant's ability to produce nuts.



Fig. 2. Late leaf spot

- The binomial name of the LLS is *phaeoisariopsis personatum*. Mycelium and haustoria are produced by the fungus. The symptoms of LLS are defined as Conidia are cylindrical or obclavate, short, hyaline which are in high humidity for three days, if the temperature is as low as 20 degrees Celsius with droplets on the lower surface, heavy dosage of fertilizers like nitrogen, phosphorus, and magnesium deficiency in soil.

### D. Rosette Diseased Leaf (RoDL)

- Fig. 3. depicted the RoDL are affected with the viruses types and lead with stunted growth and reduced yields in groundnut crops. These diseases are spread by aphids and can be controlled by using insecticides. The affected plants appear as compressed clumps with tufts of small leaves appeared on the leaves. Chlorosis and mosaic mottling can be seen on the plant.



Fig. 3. Rosette leaf

- The affected plants leaf won't develop fully as they needed for the conversion of nuts from flowers, if they convert also, it won't be good seed to harvest.

### E. Rust Diseased Leaf (RuDL)

- The Fig. 4. shows the rust diseases in groundnut crops are caused by a fungus that affects the leaves of the plant. This can cause the leaves to turn yellow and eventually drop off. The fungus can also cause the nuts to become discolored and shrivelled. The scientific name of the leaf disease called rust is *Puccinia arachidis*.



Fig. 4. Rust leaf

- The disease attacks the plant's entire aerial structure. Rust leaf disease is discovered when the groundnut crops are grown till the week of 6, then this disease will get affect and won't get good maturity of the pods. On the lower surface of the leaves, little brown to uredosori will appear. The above disease can only infect the leaf if the following circumstances are met: the humidity value above 95% high, significant rainfall, and a low temperature (20-25 degrees Celsius).

### F. Healthy Leaf (HL)

- A healthy leaf in a groundnut crop is one that is free from disease, pests, and deformities. It should be a deep green color and be able to photosynthesize efficiently. The below Fig. 5. represents the healthy leaf of groundnut crop. The peanut, scientifically known as *arachis hypogaea*, is an herbaceous annual plant that belongs to the Fabaceae family and is cultivated for its oil and edible nuts.

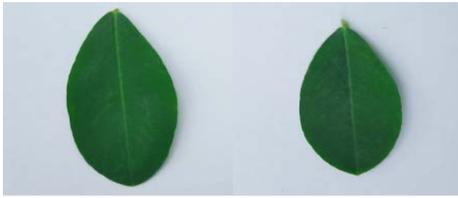


Fig. 5. Healthy groundnut leaf

- Plants that produce peanuts are typically rather tiny and have slender stems and leaves that resemble feathers. The leaves are attached to the stalk in a manner that resembles a leaf and is organized in alternate pairs. The peanut plant can produce yellow, orange, cream, or white flowers. These blooms give rise to 'pegs', which are distinctive floral structures that are pushed down into the soil to facilitate the growth of pods.

#### IV. FLOW PROCESS OF DATASET

The Fig. 6. provides a visual representation of the comprehensive real-time detection process. The first step in developing the GLDD is to collect images of both diseased and healthy groundnuts from an actual groundnut field. The initial GLDD is then subjected to several data augmentation procedures, during which it is manually annotated and expanded. The dataset was categorized into is split into two parts: healthy and unhealthy. In these two classes we have specified for training as 80% and testing as 20%. The findings of the detection include information on healthy and unhealthy classes. Groundnut Dataset Images were collected with the VIVO 2029 smartphone camera having 13 Mpx resolution to identify the groundnut crop leaves which is affected by diseases or not. This paper contains two categories healthy and diseased. In the diseased category we collected four types of leaves which are as follows with the number of leaves collected.

##### A. Dataset

Most of the dataset used for model building are taken from publicly available dataset. The study [21] also uses both the self-collected dataset and PlantVillage dataset for pepper leaf disease detection. Groundnut leaves are collected from the various fields located near the village of Pudukkottai district and also from various websites through google search engine. We have collected the five categories of leaves which is segregated with healthy and unhealthy leaves of groundnut crop. Groundnut leaf disease patterns change with the season and with other elements like humidity, temperature, a lack of NPK, and wetness.

Table II details about the dataset with 619 images of unhealthy and healthy groundnut leaves obtained, which correspond to five categories: Healthy Groundnut Leaf, ELS, LLS, rosette, and rust. These are the five leaf diseases which occurred commonly on the groundnut leaves that were chosen for two reasons. First, these common diseases can be seen with the help of images which were collected by camera itself. Furthermore, they are responsible for its significant yield decrease in groundnut crop cultivation. The above five categories of groundnut leaves are the most common diseases affected in the fields where we collected the images.

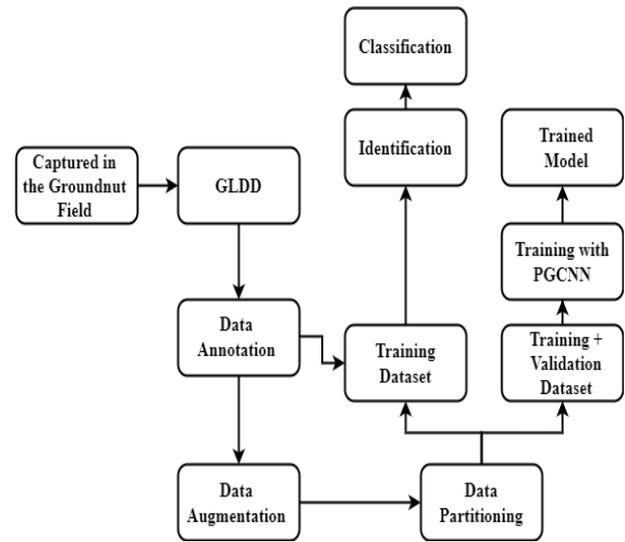


Fig. 6. Process flow of real-time detection of groundnut leaf diseases

TABLE II. DESCRIPTION OF GROUNDNUT DATASET

Sl. No	Name of the Diseases	Total No. of Images
1.	Healthy	190
2.	Unhealthy	429
<b>Total</b>		<b>619</b>

##### B. Progressive Groundnut Convolutional Neural Network (PGCNN)

1) *Input layer*: In this paper, the image size of 1380x1380x3 is given to the input layer of our model to detect whether the image is healthy or diseased. Then the images will get rescaled to 224x224x3 during the model training and testing to reduce the computational time. Here we are feeding single images of four different types of diseases with one healthy leaf category of images with no background to convolve the images with filters to identify the patterns of leaf images.

##### 2) Convolution Layer (Convnet + ReLU)

- From various types of convolution layers available in Keras API, we selected the Conv2D layer for this PGCNN model development because Conv2D layers are implemented over the spatial convolution over images. Arguments used for the Conv2D method are filters, kernel size, strides, padding, activation, and kernel initializers.
- Rescaled images of input image size 224x224x3 are convolved with 64 filters of size 3x3 kernel with strides value of (2, 2). The activation function used to take the value of feature map with "ReLU" as activation function for positive and negative values. While in convolution operation, various filters are used to detect the edges, patterns, etc., for extracting the features from the input images.
- In this paper we have used five conv2D layer to get good feature map. Operation performed by the

convolution layer uses the following equation in default, which had number of filters used in the convolution layer, filter size, padding and stride to convolve through the input image.

$$C_{out} = \frac{C_{in} + 2P - F}{S} + 1 \quad (1)$$

- $C_{out}$  represents the values which we can feed into next layer, which the input of image size 224x224x3 had been fed into the layer.  $C_{in}$  insists the size of the input image in convolution layer,  $P$  represents padding,  $F$  denotes the number of filters used in the convolution layer.

3) *Pooling*: The process of reducing the spatial dimensions (i.e., height and width of the input image) for making the computation faster by decreasing the training parameters. The pool size of the MaxPool2D is taken as (2, 2). There are 3 kinds of pooling to down-samples the input in pooling operations, they are Max Pooling, Average pooling, and Global Pooling. In this research paper, we use maxpooling2D operation to down-samples the input size of the groundnut leaf images. Max pooling chooses the most element from the feature map's filtered region. After max-pooling, the output is a feature map with the most prominent features of the preceding map.

4) *Fully Connected Layer (FC)*: The final Pooling or Convolutional Layers output is flattened to convert the matrix form into one-dimension vector and those are fed into the input of the fully connected layer. Fig. 7 explains fully connected layer is the layer which applies the linear transformation and has the input connection to every neuron into the output layer of another layer.

$$Z_{ij}(y) = f \sum_{k=1}^{n_H} w_{ij}y_k + w_{io} \quad (2)$$

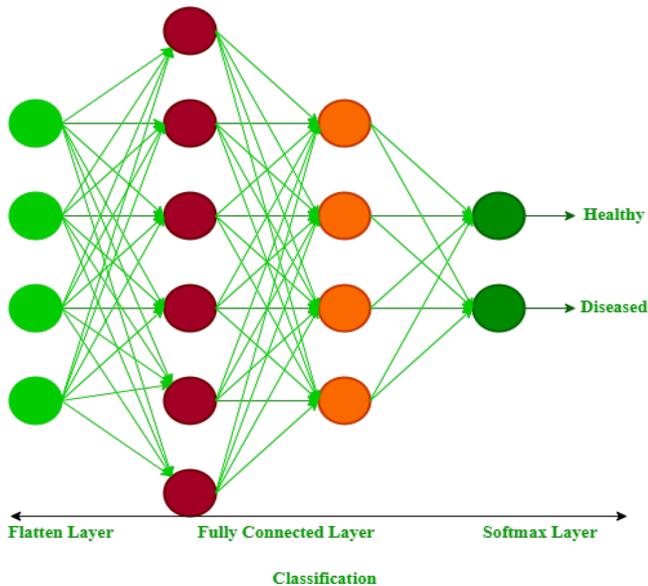


Fig. 7. FC for classification of groundnut leaf

The fully connected input layer (i.e., called flatten) is the conversion process from 2-dimensional array into a single dimensions array or linear vector. The output of the single vector is fed as the input for fully connected layer.

5) *Dense (or) output layer*: Leaky ReLU, Sigmoid, Tanh, ReLU, Softmax, and other activation functions can be used in the final layer for the output layer based on the research problem. This paper focuses primarily on the healthy or diseased category. Because we developed the model for binary classification, the sigmoid is the activation function used at the final layer of the Progressive Groundnut Convolutional Neural Network (PGCNN) to determine whether the leaf is healthy or diseased.

6) *Dropout layer*: This is one of the main characteristics of CNN. This layer is used to set the input value as 0 to skip some of the unwanted neurons to continue to next layer and also it prevents overfitting problem. Here we used 30% of neurons to be dropped out for fed the neurons into next level.

## V. IMPLEMENTATION DETAILS

### A. Groundnut Leaf Diseases Identification using PGCNN

Leaf diseases is a significant threat to agriculture, but we are facing the difficulty for the identification and classification of leaf diseases due to the lack of infrastructure development. To minimize the infrastructure cost and maximize the yield of groundnut cultivation artificial intelligence technology has evolved into farming methods also. DL is one of the subsection of Artificial Intelligence (AI), is causing a revolution in agriculture by replacing traditional methods with more efficient methods that assist farmers in the identification of leaf diseases with convolutional neural networks. The proposed PGCNN model experiments the leaf disease identification with two classes called healthy and diseased with five kinds of validation of the self-collected dataset as;

TABLE III. VALIDATION TEST WITH VARIOUS RATIO PERCENTAGES OF SELF-COLLECTED DATASET

Sl. No.	Dataset Validation %	Training Accuracy	Training Loss	Validation Accuracy	Validation Loss
1.	90-10%	98.74	03.99	91.94	0.2692
2.	<b>80-20%</b>	<b>99.19</b>	<b>03.23</b>	<b>91.94</b>	<b>0.3227</b>
3.	75-25%	98.70	03.49	92.90	0.2855
4.	70-30%	97.69	05.72	93.01	0.1745
5.	60-40%	97.30	08.77	90.32	0.4919

The Table III and Fig. 8 shows the various ranges of dataset splitting for training and testing (Validation) accuracy for an imbalanced dataset of groundnut leaf images. After plotting the table, we conclude that the dataset separation with ranges of 80% training and 20% testing works well and achieves a high of 99.19% training accuracy and 91.94% of testing accuracy than the other splitting.

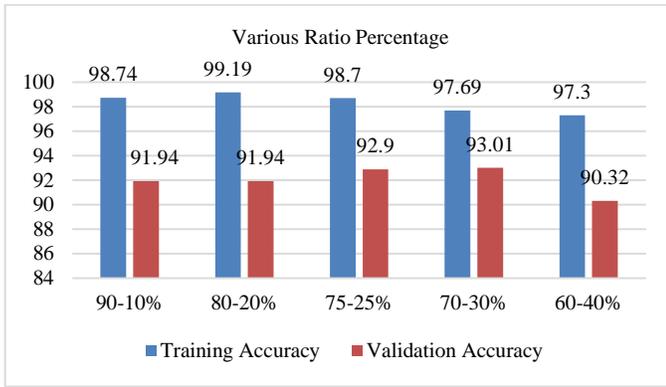


Fig. 8. Categories of ratio percentage

**B. Optimizers of PGCNN**

We have implemented five types of optimizers to reduce the loss value for our dataset which is self-prepared and validated with an 80% - 20% proportion. RMSprop, Adam, Adamax, Adagrad, and SGD were selected because these are similar to Adam optimizers with momentum.

TABLE IV. VARIOUS TYPES OF OPTIMIZERS TESTED IN PGCNN MODEL

Sl. No.	Optimizers	Training Accuracy	Validation Accuracy
1.	RMSprop	97.37	95.97
2.	<b>Adam</b>	<b>98.17</b>	<b>97.58</b>
3.	Adamax	96.56	91.13
4.	Adagrad	71.86	68.55
5.	SGD	78.34	76.61

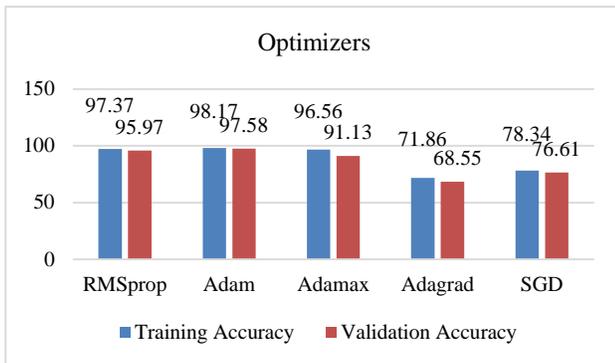


Fig. 9. Various optimizers

From the above Table V and Fig. 9 we achieved high accuracy in the Adam optimizer for training and validation with all the optimizers. The algorithm used for optimizing the PGCNN model is Adam achieves the good results in faster manner. Adam is named as adaptive moment estimation to converge the model in smooth way. This optimizer utilizes the two optimizer benefits, they are AdaGrad and RMSProp.

**C. PGCNN Model Flow Diagram**

The Fig. 10 shows the flow diagram of PGCNN model with sequence of layers which executes the model to achieve good performance. First the input image with the size of 224x224x3 has been fed into the classifier called sequential to

arrange the layers stacked one by one to form a entire network.

There are five convolutional layers constructed to convolve the image to extract the feature map as 16, 32, 64, 128 etc. First Conv2D layer with 64 filter to specify the depth of the filters and 3x3 dimensions of kernel/filter represents the height and depth of the matrix with padding same value, following the activation function as “ReLU” of non-linear function. Then to reduce the dimensionality of images we apply max\_pooling of 2x2 function. This reduces the size of the images with half the value from the previous layer input. By this way we are repeating five times of the convolution operation with filter of 64 on 3x3 size kernel by applying “relu” activation function with the conclusion of convolution is max\_pooling layer.

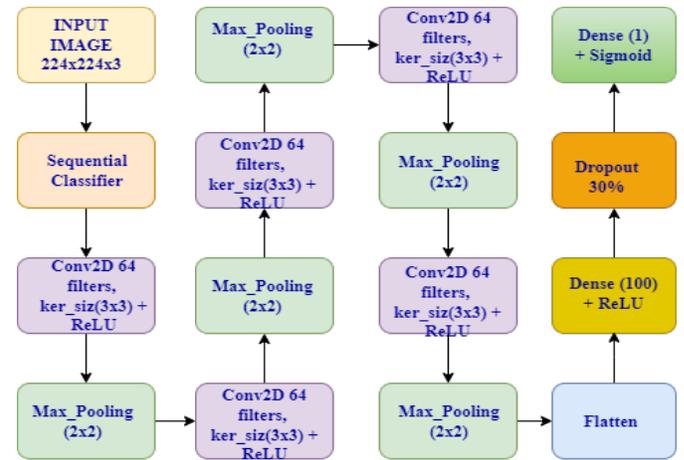


Fig. 10. PGCNN model flow diagram

TABLE V. VARIOUS RATIO OF PGCNN MODEL FOR TRAINING AND VALIDATION ACCURACY

Sl. No.	Sample Images	Training Ratio (%)	Validation Ratio (%)	Training Accuracy	Validation Accuracy
1	619	90	10	0.9910	0.9677
2		<b>80</b>	<b>20</b>	<b>0.9939</b>	<b>0.9758</b>
3		70	30	0.9954	0.9140
4		60	40	1.0000	0.8992

After completing the convolution operation we are applying the flatten() class to convert the two dimension image into one dimensional vector. With 100 nodes in our output layer we are having the dense layer as the previous layer of output layer to spread the probability of values from the fully connected layer by the activation function as “ReLU”. To improve the performance of PGCNN model and reduce the overfitting we added dropout function with 30% neurons to be dropped out randomly by selecting the important neuron and some of the neurons are ignored during training process. Dropout() is one of the type of regularization in deep learning models.

Finally, we concluded with dense layer with single neuron for binary classification and the activation function of “sigmoid” is used for the classification of PGCNN model as healthy or unhealthy images. We used sigmoid activation function because this function supports the binary classification for PGCNN model.

## VI. RESULTS

The model creation and executions were performed in TensorFlow framework. Dataset has been prepared by self-collected with the samples of total 619 groundnut leaves as two categories such as Unhealthy and Healthy options. The leaves were collected from the village near by Pudukkottai town, Tamilnadu at different times with different groundnut crops.

The Table VI, Fig.11 and Fig. 12 explains about the results for the various ratios for the proposed PGCNN model for the groundnut leaves disease identification and classification of healthy and unhealthy category. We have taken four different ratios as 90-10%, 80-20%, 70-30% and 60-40% respectively. From these compilations we have achieved the good training accuracy and validation accuracy in 80-20% ratio. So, we freeze that ratio for further evaluation for PGCNN model. With the minimum number of sample images, the proposed model achieved 99.39% as the accuracy of training the model and 97.58% as the accuracy of validation of PGCNN model. The 90-10% ratio also achieves good training accuracy as 99.10% and validation accuracy as 96.77% with the training loss of 0.0104%, validation loss as 18.79%. Next ratio tested with 70-30%, it achieves the training accuracy of 99.54% and validation accuracy of 91.40% with 0.0111% of training loss and 90.05% of validation loss. At last we have tested with the ratio of 60-40% which achieves the training accuracy of 100% and 89.92% of validation accuracy with the training loss of 0.0041% and 39.59% validation loss.

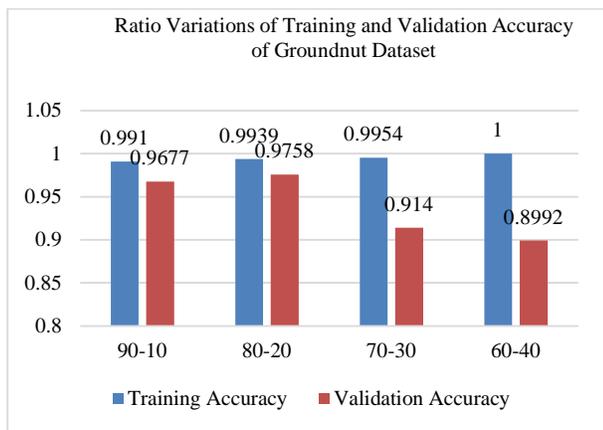


Fig. 11. Model training and validation accuracy graph

TABLE VI. VARIOUS RATIO OF PGCNN MODEL FOR TRAINING AND VALIDATION LOSS

Sl. No	Sample Images	Training Ratio (%)	Validation Ratio (%)	Training Loss	Validation Loss
1.	619	90	10	0.0194	0.1979
2.		80	20	0.0148	0.1806
3.		70	30	0.0111	0.9005
4.		60	40	0.0041	0.3959

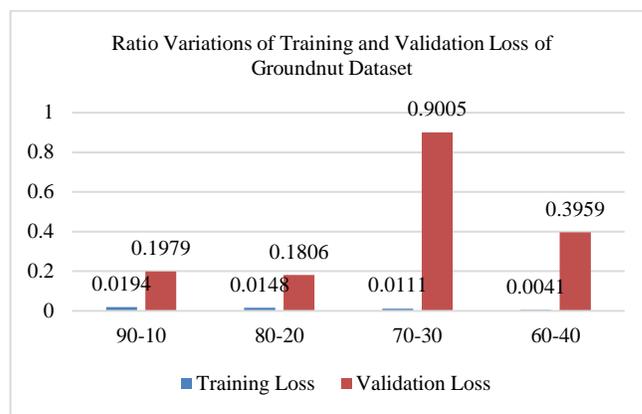


Fig. 12. Model training and validation loss

From the above details we have frozen the ratio for our PGCNN model as 80-20% for further evaluation of predicting the image as healthy or unhealthy. Because it ended up with good accuracy on both training and validation with less training and validation loss when compared to the other ratio results.

The Fig. 13. explains that the identification of groundnut leaves as healthy or unhealthy by the classes we defined in the dataset, it classifies the category by class names if it falls on healthy or unhealthy classes. And also it shows that the label defined as 1 for unhealthy class and 0 for healthy class.

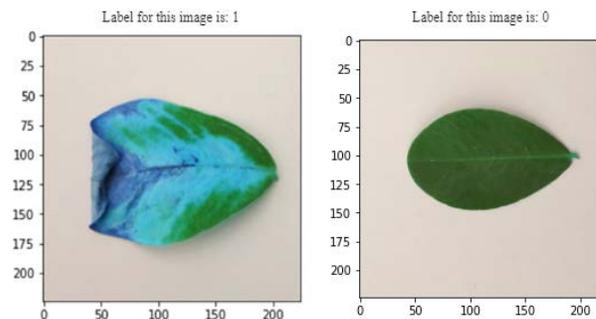
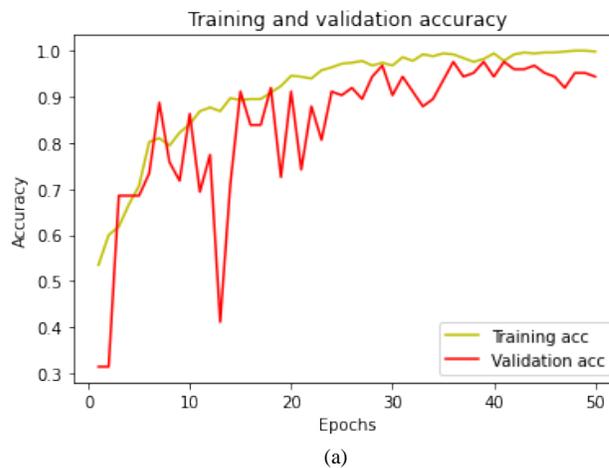


Fig. 13. Identification and classification of groundnut leaf as unhealthy or healthy



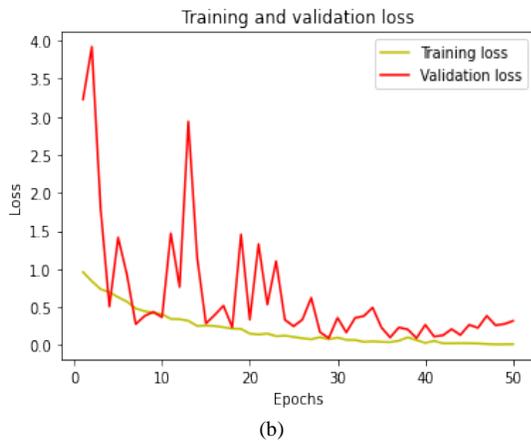


Fig. 14. (a). PGCNN model training and validation accuracy, (b). PGCNN model training and validation loss

The Fig. 14(a) and Fig. 14(b) shows that the graph for the proposed model of PGCNN training, validation of accuracy and loss against the groundnut dataset which was collected from the field located nearby the Pudukkottai district. The curve indicates how the model fits into the accuracy of both training and validation data. The training loss indicates that the training data learns well and it decreases over time to fit the curve as good fit. The validation loss jumps up and down to fit into the curve for model that learns the new data to fit the curve in good fit. The above model learns all data in a good manner to fit the curve of loss that decreases in certain time. Likewise, the training and validation accuracy curve also fits in good manner by gradual increase in the curve.

TABLE VII. COMPARATIVE ANALYSIS

Sl. No	CNN Models	Training Accuracy	Training Loss	Validation Accuracy	Validation Loss
1	PGCNN	0.9939	0.0194	0.9758	0.1806
2	AlexNet	0.9737	0.1039	0.7419	0.1148
3	VGG11	0.6949	0.6160	0.6855	0.6229
4	VGG13	0.6949	0.6159	0.6855	0.6228
5	VGG16	0.6949	0.6161	0.6229	0.6855
6	VGG19	0.6949	0.6161	0.6855	0.6229

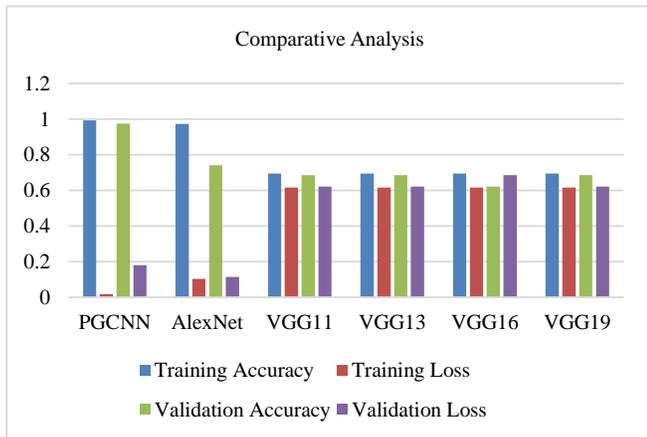


Fig. 15. Various model comparison on proposed PGCNN model

The Table VII and Fig. 15 shows that the comparative analysis of six CNN architecture including the proposed PGCNN model. It shows that PGCNN gives good accuracy and reduced loss for training and validation process. Here we have taken AlexNet, VGG11, 13, 16 and 19 models. The reason behind the model selection like VGG series is that we have built the model based on this architecture. By comparing the obtained results of proposed model with the pretrained models like VGG architectures. The proposed model has achieved the good results as compared to VGG model architecture.

## VII. DISCUSSION

This research paper discusses about the groundnut leaf diseases identification and classification as whether the leaf is healthy or unhealthy. For experimenting this dataset with appropriate base model we have gone through various paper as related works to concluded our model. With the use of literature review we confirmed to develop PGCNN model which we used CNN architecture as our baseline to implement the model. After implementing the model, we have compared with the pretrained CNN architecture for performance comparisons. We achieved the good performance accuracy compared to the pretrained models which has limited layers. The proposed model achieved the training accuracy of 99.39% and validation accuracy of 97.58%. Such that the training loss as 0.01% and validation loss as 0.18% respectively. These results achieved with 619 total images for the dataset which was collected from the fields. In future we are planning to collect more images for large dataset and also, we will be preparing our dataset to make as a benchmark dataset for publicly available for the researcher to use and explore it. In next section we concluded our research work that have done with the dataset by PGCNN model.

## VIII. CONCLUSION

This paper was implemented with the dataset which was collected from the real-life scenario of field located nearby Pudukkottai, Tamil Nadu, India. Groundnut dataset has two category of groundnut leaves namely healthy and unhealthy for identification of diseases from the leaves. The detailed description of the groundnut dataset has been listed under the Section IV. Here we focused to identify and classify the dataset with two classes. Five different CNN architecture were designed and executed to compare the proposed model whether the PGCNN model performs well when compared to the pretrained architectures like VGG11, 13,16, etc. we used performance metrics as accuracy for all the architectures, in that PGCNN achieves good accuracy for both training and validation process. The proposed PGCNN training accuracy was 99.39% and validation accuracy was 97.58%. The model loss also reduced as compared to other models as training loss 0.0194 and validation loss as 0.1806.

In future we will be focusing the dataset on multi-class classification to classify the leaves by their diseased names and incorporate it into the prediction-based solutions. We will be focusing on more data to be collected on various seasons and measures the performance.

REFERENCES

- [1] R. M. Rakholia, J. H. Tailor, J. R. Saini, J. Kaur, and H. Pahuja, "Groundnuts Leaf Disease Recognition using Neural Network with Progressive Resizing," *Int. J. Adv. Comput. Sci. Appl.*, vol. 13, no. 6, pp. 83–88, 2022, doi: 10.14569/IJACSA.2022.0130611.
- [2] S. S. Koshy, V. S. Sunnam, P. Rajgarhia, K. Chinnusamy, D. P. Ravulapalli, and S. Chunduri, "Application of the internet of things (IoT) for smart farming: a case study on groundnut and castor pest and disease forewarning," *CSI Trans. ICT*, vol. 6, no. 3–4, pp. 311–318, 2018, doi: 10.1007/s40012-018-0213-0.
- [3] M. Ramakrishnan and A. N. A. Sahaya, "Groundnut leaf disease detection and classification by using back propagation algorithm," 2015 Int. Conf. Commun. Signal Process. ICCSP 2015, no. 7092512506, pp. 964–968, 2015, doi: 10.1109/ICCSP.2015.7322641.
- [4] M. Lv, G. Zhou, M. He, A. Chen, W. Zhang, and Y. Hu, "Maize Leaf Disease Identification Based on Feature Enhancement and DMS-Robust Alexnet," *IEEE Access*, vol. 8, pp. 57952–57966, 2020, doi: 10.1109/ACCESS.2020.2982443.
- [5] P. Jiang, Y. Chen, B. Liu, D. He, and C. Liang, "Real-Time Detection of Apple Leaf Diseases Using Deep Learning Approach Based on Improved Convolutional Neural Networks," *IEEE Access*, vol. 7, pp. 59069–59080, 2019, doi: 10.1109/ACCESS.2019.2914929.
- [6] B. Liu, C. Tan, S. Li, J. He, and H. Wang, "A Data Augmentation Method Based on Generative Adversarial Networks for Grape Leaf Disease Identification," *IEEE Access*, vol. 8, pp. 102188–102198, 2020, doi: 10.1109/ACCESS.2020.2998839.
- [7] C. Zhou, Z. Zhang, S. Zhou, J. Xing, Q. Wu, and J. Song, "Grape leaf spot identification under limited samples by fine grained-GAN," *IEEE Access*, vol. 9, pp. 100480–100489, 2021, doi: 10.1109/ACCESS.2021.3097050.
- [8] S. B. Jadhav, V. R. Udipi, and S. B. Patil, "Identification of plant diseases using convolutional neural networks," *Int. J. Inf. Technol.*, vol. 13, no. 6, pp. 2461–2470, 2021, doi: 10.1007/s41870-020-00437-5.
- [9] C. Zhou, S. Zhou, J. Xing, and J. Song, "Tomato Leaf Disease Identification by Restructured Deep Residual Dense Network," *IEEE Access*, vol. 9, pp. 28822–28831, 2021, doi: 10.1109/ACCESS.2021.3058947.
- [10] Q. Wu, Y. Chen, and J. Meng, "Dcgan-based data augmentation for tomato leaf disease identification," *IEEE Access*, vol. 8, pp. 98716–98728, 2020, doi: 10.1109/ACCESS.2020.2997001.
- [11] R. R. Atole and D. Park, "A Multiclass Deep Convolutional Neural Network Classifier for Detection of Common Rice Plant Anomalies," no. February, 2018, doi: 10.14569/IJACSA.2018.090109.
- [12] P. Kaur, S. Harnal, V. Gautam, M. P. Singh, and S. P. Singh, "A novel transfer deep learning method for detection and classification of plant leaf disease," *J. Ambient Intell. Humaniz. Comput.*, no. 0123456789, 2022, doi: 10.1007/s12652-022-04331-9.
- [13] Suresh and K. Seetharaman, "Real-time automatic detection and classification of groundnut leaf disease using hybrid machine learning techniques," *Multimed. Tools Appl.*, 2022, doi: 10.1007/s11042-022-12893-1.
- [14] H. Amin, A. Darwish, A. E. Hassanien, and M. Soliman, "End-to-End Deep Learning Model for Corn Leaf Disease Classification," *IEEE Access*, vol. 10, pp. 31103–31115, 2022, doi: 10.1109/ACCESS.2022.3159678.
- [15] R. Thangaraj, S. Anandamurugan, and V. K. Kaliappan, "Automated tomato leaf disease classification using transfer learning-based deep convolution neural network," *Journal of Plant Diseases and Protection*, vol. 128, no. 1, pp. 73–86, 2021, doi: 10.1007/s41348-020-00403-0.
- [16] M. P. Vaishnav, K. Suganya Devi, and P. Ganeshkumar, "Automatic method for classification of groundnut diseases using deep convolutional neural network," *Soft Comput.*, vol. 24, no. 21, pp. 16347–16360, 2020, doi: 10.1007/s00500-020-04946-0.
- [17] S. Muthukumar, P. Geetha, E. Ramaraj, and C. Head, "Leaf Spot Disease Image Classification for Groundnut Crop using Deep Convolutional Neural Network," *Turkish J. Comput. Math. Educ.*, vol. 12, no. 14, pp. 3166–3179, 2021.
- [18] K. Suganya Devi, P. Srinivasan, and S. Bandhopadhyay, "H2K – A robust and optimum approach for detection and classification of groundnut leaf diseases," *Comput. Electron. Agric.*, vol. 178, no. October, p. 105749, 2020, doi: 10.1016/j.compag.2020.105749.
- [19] M. Janani and R. Jebakumar, "Detection and classification of groundnut leaf nutrient level extraction in RGB images," *Adv. Eng. Softw.*, vol. 175, no. October 2022, p. 103320, 2023, doi: 10.1016/j.advengsoft.2022.103320.
- [20] R. Dwivedi, T. Dutta, and Y. C. Hu, "A Leaf Disease Detection Mechanism Based on L1-Norm Minimization Extreme Learning Machine," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, doi: 10.1109/LGRS.2021.3110287.
- [21] M. Ahmad, M. Abdullah, H. Moon, and D. Han, "Plant Disease Detection in Imbalanced Datasets Using Efficient Convolutional Neural Networks with Stepwise Transfer Learning," *IEEE Access*, vol. 9, pp. 140565–140580, 2021, doi: 10.1109/ACCESS.2021.3119655.

# Enhancing Image for CNN-based Diagnostic of Pediatric Pneumonia through Chest Radiographs

Vaishali Arya, Tapas Kumar

Department of CSE, FET, Manav Rachna International Institute of Research and Studies, Faridabad, India

**Abstract**—In underdeveloped nations, severe lower respiratory infections are the principal reasons of infant mortality. The best treatments and early diagnosis are now being used to alleviate this issue. In developing nations, better treatment and prevention approaches are still required. Clinical, microbial, and radiographic clinical studies have a broad range of applicability within and across populations, and it much depends on the knowledge and resources that are made accessible in different situations. The most appropriate procedure is a chest radiograph (CXR), although pediatric chest X-ray techniques using machine intelligence are uncommon. A strong system is required to diagnose pediatric pneumonia. Authors provide a computer-aided diagnosis plan for the chest X-ray scans to address this. This investigation provides a deep learning-based intelligent healthcare that can reliably diagnose pediatric pneumonia. In order to improve the appearance of CXR pictures, the suggested technique also employs white balancing accompanied with contrast enhancement as a preliminary step. With an AUC of 99.1 on the testing dataset, the suggested approach outscored other state-of-the-art approaches and produced impressive results. Additionally, the suggested approach correctly classified chest X-ray scans as normal and pediatric pneumonia with a classification accuracy of 98.4%.

**Keywords**—Contrast enhancement; convolution neural network; pediatric pneumonia; radiography; white balancing

## I. INTRODUCTION

The most prevalent pathogens being the reason of mortality in kids is pneumonia. Pneumonia is a disease caused by infections such as viruses, bacteria, and fungi that enter the lungs and cause the alveoli to fill with inflammatory fluid. It's crucial to get this condition diagnosed early since it has devastating consequences, particularly in kids under the age of five, and can be controlled by doing so. More than 150 million kids below the age of five are affected with pneumonia annually, and 20 million of kids require hospitalization for treatment [1]. The most of pneumonia infections occur in developing and underdeveloped countries, where there is a lack of medical services, excessive urbanisation, pollution, and unclean air. Therefore, avoiding the condition from turning deadly can be greatly aided by early treatment and detection.

The conventional biological diagnostics are inadequate to determine the cause of pediatric pneumonia since blood cultures are insensitive, pulmonary aspirates are difficult to collect, and antigen assays have poor specificity [2]. Presently, the diagnosis of pneumonia is determined by the patient's symptoms, the results of a CXR, the growth and susceptibility of the bacteria found in throat swabs or sputum specimens, and blood tests. Early detection of pediatric pneumonia is crucial in

reducing repercussions since this illness is curable and may be prevented, particularly through vaccines. CXRs, while having a lower resolution than Magnetic Resonance Imaging (MRI) or Computerized Tomography (CT) scans, can nevertheless be utilized to conduct a wide range of evaluations. Radiographic diagnosis of pediatric pneumonia is very subjective and is reliant on the radiologist's skill and understanding. High definition MRI and CT scans make it simpler to detect pneumonia, though most radiologists prefer to utilize CXRs for evaluations due to the faster turnaround time and economical nature of the technology. Radio-opacities or white patches in the airways, especially in the alveoli, are typically seen on a scan of pneumonia and signify the existence of chronic effusion. Attributed to the reason that some illnesses might resemble similar indications, these radiological observations can be challenging for a trainee radiologist and result in incorrect positives and negatives. Fig. 1 displays examples of CXR pictures from the pediatric population that have been categorized as normal and pneumonia for this research.

Artificial intelligence (AI) has notably been utilized to diagnose pneumonia-related abnormalities in radiographic scans. Deep Learning (DL) techniques are the most well-known and commonly used approach for detecting and classifying clinical pictures in general, and pneumonia specifically, owing to the abundance of tagged CXR resources and ubiquitous, reasonably priced computer capacity. These technologies carry the potential and can diagnose numerous illnesses with conventional physician precision [3].

In earlier investigations, various unique data pre-processing methods have been demonstrated to be helpful in a wide range of applications [4], including voice recognition, hand motion detection utilizing sonography [5], and hand motion identification utilizing navigation systems [6]. In the job of classifying natural photographs, Convolutional Neural Networks (CNNs) has demonstrated excellent outcomes. Clinical scans, on the other hand, can be found in three-dimensional RGB, four-dimensional, or two-dimensional with grayscale files, whilst natural photographs are often found in two dimensional RGB file types. This necessitates a lot of adjusting. Another important distinction relates to illumination changes. In addition, natural images are typically recognised by their edges, fundamental shapes, relationships between nearby pixels, etc., while each pixel's illuminance extent or severity are not important characteristics for their identification. Conversely, when it comes to clinical scans, particularly CXR scans, each pixel's luminance is key for identifying the image's impacted areas. This is incredibly important for spotting anomalies in CXR scans. So, developing

a reliable and effective model for spotting anomalies in clinical scans requires good image pre-processing. Thus, current study adopts a technique that combines white balance and contrast enhancement as a stage in the pre-processing of input for picture improvement.

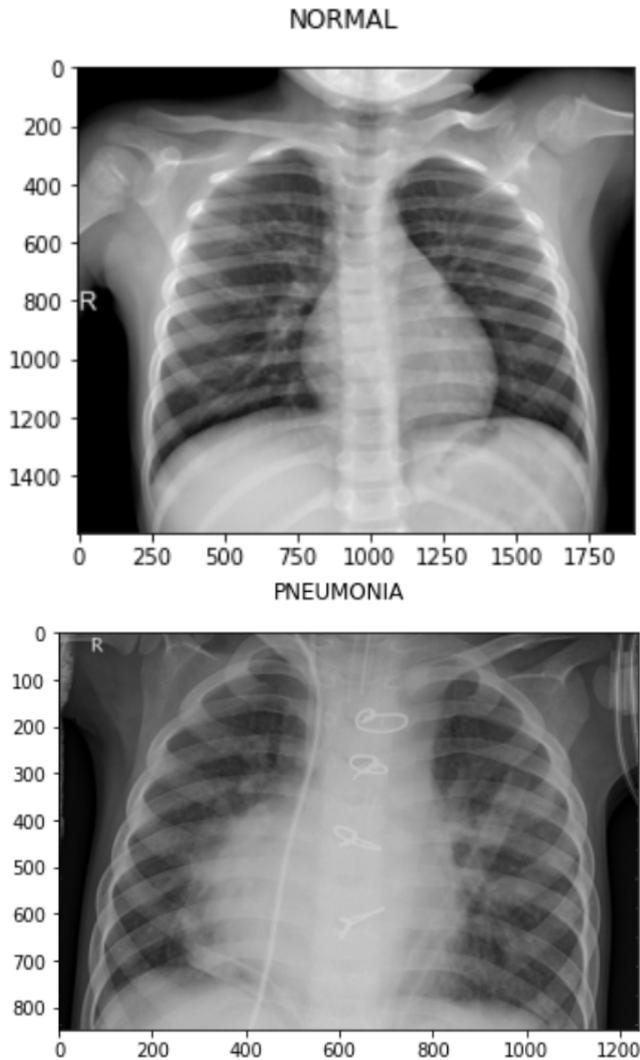


Fig. 1. Samples of CXR images.

The system takes advantage of a multi-layered CNN to accurately predict pneumonia by automatically extracting information from radiography scans. While screening pediatric pneumonia, computer assisted diagnostic (CAD) technologies can take the place of radiologist discretion. They are also useful in rural places or nations without access to technology, such as radiological competence, to corroborate diagnostic evidence. This study suggests a cutting-edge model that can distinguish between pneumonia and the usual. The key contributions of this research are listed below.

- With a remarkable good accuracy of 98.4%, the new proposed design can differentiate between pediatric pneumonia and normal.

- In addition to accuracy, the model's sensitivity, specificity, precision, and F1-score are all 0.9841, 0.9837, 0.9898, and 0.9867, respectively.
- The error symbolized by the False Positives (FP) and False Negatives (FN) is close to 1.6%.
- The efficiency of the suggested design is significantly better than the most recent models stated in the literature.

The manuscript is further divided into sections. Section II highlights the related work. Section III introduces the proposed model. Section IV elaborates the experiment conducted and results achieved. Section V, finally concludes this research.

## II. RELATED WORKS

Major innovations in deep learning have made it possible to automate CXR interpretation with accuracy on par with that of professional specialists [7, 8, 9, 10]. By decreasing ambiguities in explanation and encouraging increasingly broad use of radiological outcomes in scientific studies, automating the diagnosis through CXR may increase sample performance.

Transfer learning was used by Kermany et al. [11] to build a CNN model that could identify pneumonia in CXR scans. CXR images have been classified by Rajaraman et al. [12] as diagnose bacterial and viral pneumonia using a CNN-based approach. Rather than employing the entire picture, they trained CNN models using regions of concern that only contain the lungs. A 121-layer CNN model called CheXNet was created by Rajpurkar et al. [13]. One million CXR scans with 14 distinct illnesses were used to train CheXNet. 420 CXR scans have been used to evaluate the suggested model, and the findings were contrasted to those of professional radiologists. As a consequence, it has been discovered that the CNN model outperformed radiologists generally in spotting pneumonia. A CNN model was presented by Stephen et al. [14]. They built the CNN model from the beginning to classify images from a particular CXR scan to obtain outstanding generalization ability that they utilised to assess whether or not a patient has been afflicted with pneumonia. This differs from prior methods relying simply on transfer learning or classic handmade approaches.

Liang et al. [15] used backpropagation and expanded convolution techniques to achieve pneumonia diagnosis with a CNN framework. While categorising CXR scans, they additionally identified the transfer learning impact on convolutional networks. A sequential CNN model comprising eighteen layers that automatically diagnoses pneumonia has been suggested by Siddiqi [16]. Gu et al. [17] suggested a two-stage technique for discriminating between distinct types of pneumonia. The transfer learning approach has been employed by Rahman et al. [18] to conduct the pneumonia prediction leveraging four pre-trained CNN models. They used three distinct categorization techniques to label CXR scans. Three well-known CNN models have been employed by Togacar et al. [19] for the feature extraction step. Researchers built every model independently employing the same input, and based on the final fully linked layers of every CNN, they extracted 1000 characteristics. For the pneumonia identification challenge,

1000 characteristics have been produced, and these characteristics have been used as inputs to machine learning classifier. A CapsNet CNN architecture with multi-layered containers has been introduced by Mittal et al. [20] for the detection of pneumonia in CXR scans.

### III. PROPOSED MODEL

The objective of current research is to devise an automatic pediatric pneumonia recognition system that can assist in the absence of radiologist. Fig. 2 shows the strategy suggested in this research for identifying pediatric pneumonia utilizing CXR scans. The whole architecture is divided into four phases, i.e., pre-processing model, pre-trained model, interim model and outcome model.

#### A. Pre-processing Model

The image pre-processing model enrich the input by performing four operations i.e., image enhancement, normalization, resizing and data augmentation.

1) *Image enhancement*: Image enhancement start with white balancing followed by contrast enhancement. The image processing technique known as white balance is used to correct the colour integrity of a CXR scan. The scan recording technology does not perfectly capture light like the naked eye does because of the low illumination circumstances in clinical scans, which made certain portions of the scan look dark. The finished scan should thus accurately reflect the colours of the actual picture through image processing or restoration. This investigation's goal is to make the scan more visible so that the suggested model may identify important information from it. Through the separate stretching of the RGB channels, the white balance technique modifies the colours of the scan's layered structure. Stretching is done for the leftover colour range whilst discarding the pixel colours that are near the terminus of the RGB channels and are only employed by

0.05% of the pixels that comprise the scan. After this procedure, the minimum and maximum limit readings would not be adversely affected by pixel colours that were seldom existent at the channel's terminus when stretching [21]. In this solution, we have implemented a white balance as depicted in Eq. (1) and Eq. (2).

$$I_p = \frac{1}{3} \left( \sum_{i \in R,G,B} \frac{1}{W * H} \sum_{x=1}^W \sum_{y=1}^H \text{image}_i(x,y) \right) * p\% \quad (1)$$

$$\text{Image}_{\text{white\_balance}} = \text{Sat} \left( \frac{\sum_{i \in R,G,B} \text{image}_i - I_p * 255}{I_{99.05} - I_{0.05}} \right) \quad (2)$$

where  $I_p$  denotes the  $p^{\text{th}}$  aggregate percent of RGB channels, and  $\text{Sat}()$  performs saturation operation within the range of (0,255).  $\text{Image}_{\text{white\_balance}}$  presents the white balanced channels pixel values.

After the white balancing, contrast enhancement is achieved through Contrast Limited Adaptive Histogram Equalization (CLAHE) [22], which is an upgradation of adaptive histogram equation (AHE) [23]. AHE divides the original picture into a number of smaller pictures, sometimes referred to as tiles. This approach employs the intensity rebinding algorithm for every tile to be derived from the histogram of every tile, which is generated and correlates to various regions of the picture.

This technique over-intensifies the image, which brings noise into the picture [23]. To handle this noise, authors adopt the CLAHE algorithm. The CLAHE algorithm operates exactly like the AHE algorithm, except before generating the continuous distribution function, it slices the histogram at particular values to restrict the intensification. The histogram's over-intensified portion is further dispersed throughout the histogram. CLAHE had remarkable results in improving CXR scans in one of the earlier research [24] and was deemed helpful in analysing a wide range of medical pictures.

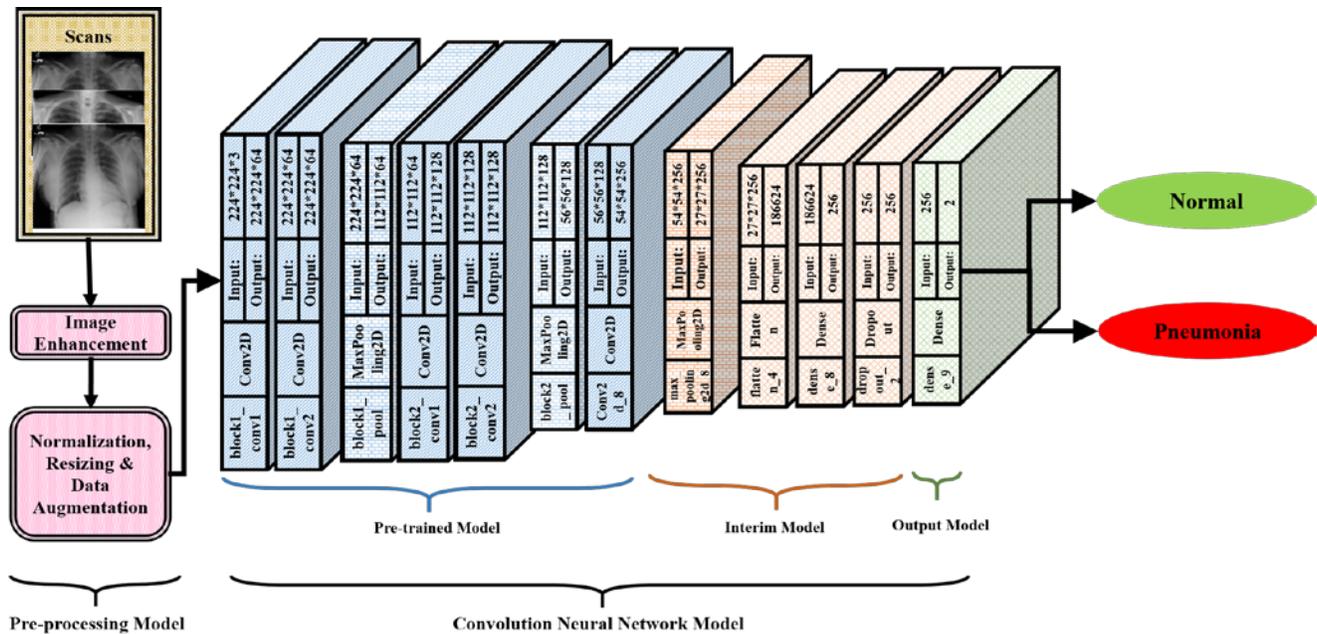


Fig. 2. Proposed model.

2) *Normalization, resizing & data augmentation*: Low-quality or low-resolution CXR scans have not been excluded; all of the CXR scans have been used, normalised, and resized to 224\*224\*3 as the current proposal specifies. A huge collection is needed for the neural network to be trained effectively. CNN models rarely generalise when built on a smaller dataset, which results in low test performance [25, 26]. One approach to resolving this issue is data augmentation, which effectively enhances the underlying dataset. There have been 3883 pneumonia-infected CXR scans and 1349 normal CXR scans in the train set employed in this research. Just the scans of the normal case have been augmented two times since the collection already provided adequate scans of the pneumonia incident. Authors augmented the CXR scans in three steps considering that not all augmentation techniques worked well with CXR images: *i*) random flipping (to negotiate directly with pneumonia signs on every edge of the CXR), *ii*) random shearing (to get a profound understanding of the relationship between pixels), and, *iii*) differing rotation [27]. After augmentation, there have been 3883 scans of pneumonia and 4047 scans of normal. Scans from the testing dataset have not been enhanced.

#### B. Convolution Neural Network Model

The proposed CNN model is divided into three phases 1) Pre-trained model, 2) Interim model, and 3) Outcome model.

1) *Pre-trained model*: The pre-processed scans are stuffed as input to the pre-trained model. This phase of the proposed model is a CNN framework with two convolution filtering layers on top and one pooling layer. Subsequently, one pooling layer and three convolution filtering layers are applied twice. The SoftMax conclusion is the result of three entirely interconnected layers that make up the architecture's intellect. One million ImageNet [7] annotated images with 224\*224\*3 colour images assigned to 1000 class labels are used to pre-train the classifiers. Authors have chosen the top seven layers of this framework to serve as our pre-trained model.

2) *Interim model*: Given the fact that our intended collection of CXR scans comprises "big 3-channeled images," authors initially make minor restructuring to the pre-trained model while keeping its core values. Through our own unique layers, we deliberately altered the final grades of the initial model. Our method employs three filters to create a three-channel extracted features while maintaining kernel size and latency hyper-parameters that are similar to the primary convolutional of the original incarnation. Authors utilize the legitimate activation function while applying the preliminary idea. Authors want to refine the previous concept for a "middleman" zone while both training and developing the layers to get the interim model. The last level of the interim phase is modified to anticipate certain classes since the large clinical images are more akin to CXR scans. This model's outcome is identical to the preceding tightly compacted layer input, with arbitrary vertices set to zero. The result would be labeled as  $Output_{interim}$ .

3) *Outcome model*: In order to connect the interim model results to the arena of CXR scan, which constitutes a last dense layer employing "softmax" activation function, authors adopt 224\*224\*3 inputs obtained from pre-trained model. Researchers can use a larger input dimension in addition to the advantage of translating a model of imaging modalities to the CXR arena. This has the benefit of producing more distinctive patterns since larger pictures contain greater info. Authors include a pre-trained model using the 3-channeled images from the result zone after 50 epochs of the interim solution. This output model distinguishes between normal and pediatric as  $Output_{final}$ .

$$Output_{final} = \max\_index \left( \frac{e^{Output_{interim_i}}}{e^{i1} + e^{i2}} \right) \quad (3)$$

where,  $Output_{interim_i}$  denotes the input vector,  $e^i$  presents the exponential function and  $\max\_index()$  operation calculates the index of class (i.e., 1 or 2) depicted by maximum instances of input vector.

#### IV. EXPERIMENTS AND RESULTS

This section provides the precise details of the tests that have been carried out to evaluate the suggested framework. The CNN model have been implemented using the Keras toolkit alongside TensorFlow. On a machine featuring 64 GB of RAM and an NVIDIA 1080 Ti graphics card, processing has been conducted.

##### A. Dataset

CXR scans of kids are much more troublesome than those of grownups. Particularly, poor posture and even the existence of limbs like arms and necks pose a challenge to educate the network and locate the pneumonia that is associated with the issue. A collection of CXR scans collected by Kermany et al. [28] has been employed in the investigation. This dataset comprises of 5856 scans of kids between the ages of one and five that have been classified as pneumonia and normal by qualified radiologists. Two subgroups of the training examples in the dataset have been provided: 1349 normal scans and 3883 pneumonia scans. Similarly, the testing dataset in the scans have been split into 234 normal scans and 390 pneumonia scans. 20% of the training examples has been employed for validation, and the remaining 80% for training. The adopted dataset has already secured the testing data. This data collection has been employed in several academic investigations, and various categorization techniques have been applied on it [29].

##### B. Results

The proposed framework has been trained to distinguish between scans of pediatric pneumonia and scans of normal CXR, and the prototype appears to have learnt how to address this issue successfully and effectively. It appears to have been successful in locating the properties associated with the particular category. Fig. 3 and Fig. 4 depict the accuracy and loss attained in the proposed model corresponding to 50 epochs, respectively. Considering accuracy and loss values, the suggested model generated the highest outcomes.

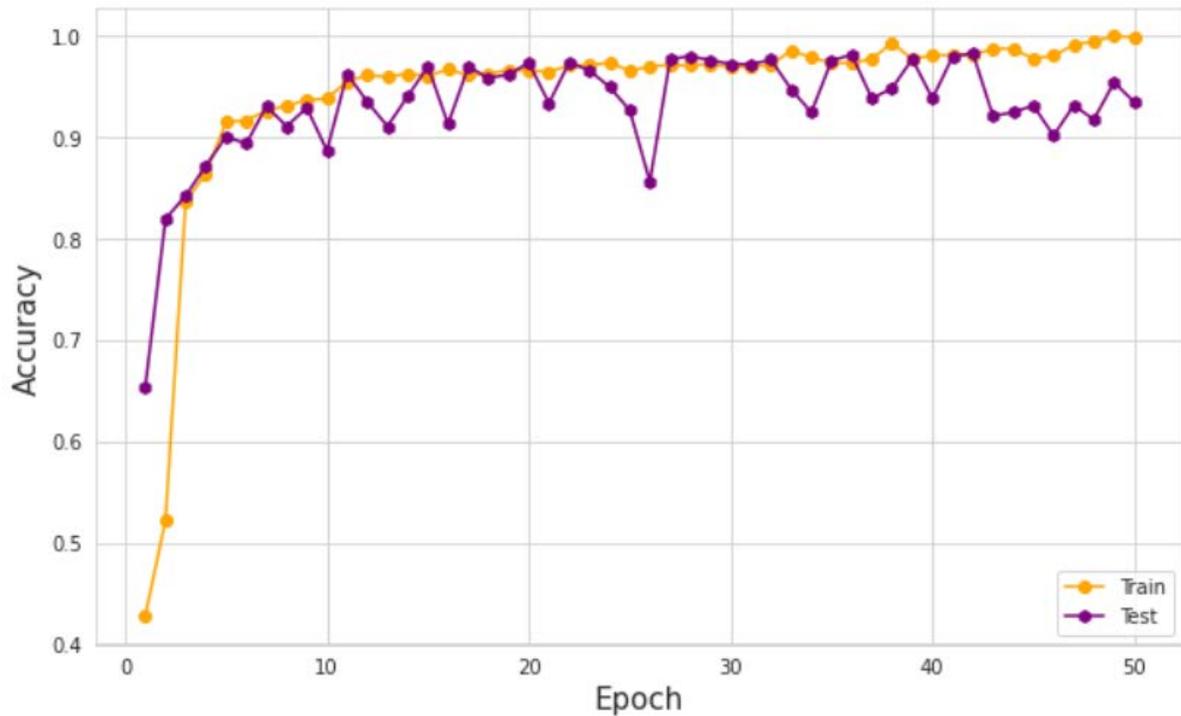


Fig. 3. Accuracy comparison of training and testing dataset.

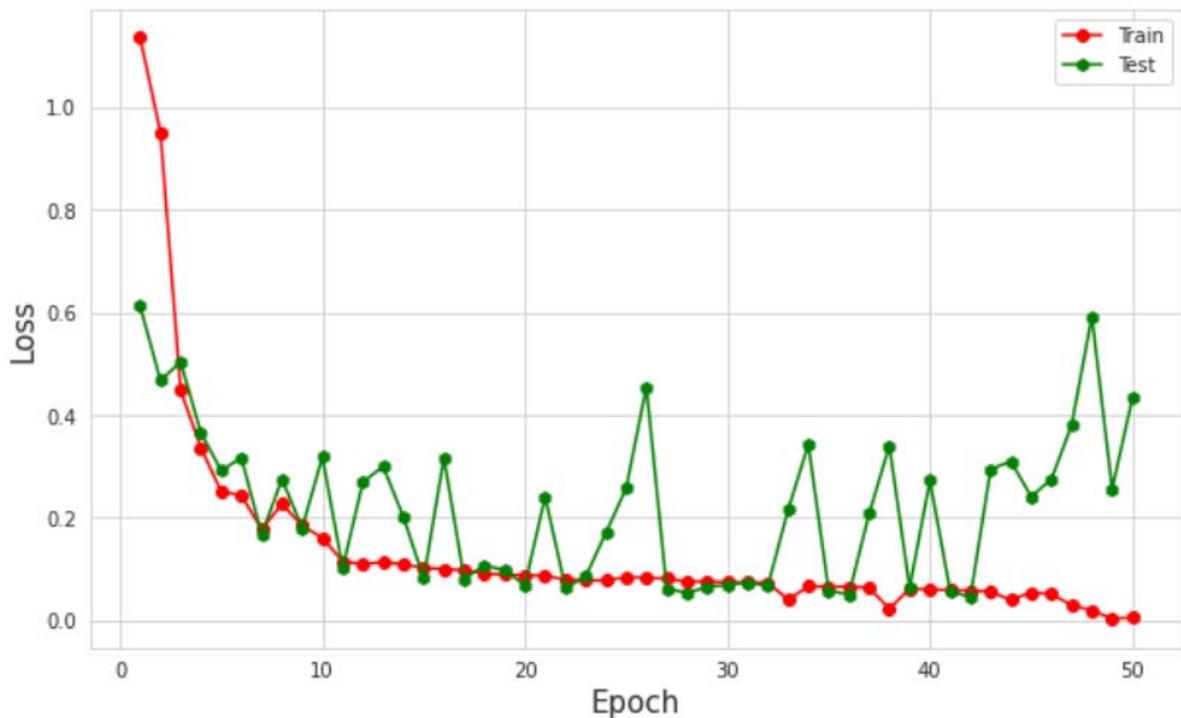


Fig. 4. Loss comparison of training and testing dataset.

The confusion matrix is depicted graphically in Fig. 5 of the classifier's effectiveness, in which the rows correspond to the forecasts and the columns to the original category. The amount of correctly and incorrectly classified scans is shown in the confusion matrix, and it is evident that the suggested framework has been able to distinguish between the underlying

categories with 98.4% accuracy. Further, the error symbolized by the False Positives (FP) and False Negatives (FN) is close to 1.6%. In addition to accuracy, the model's sensitivity, specificity, precision, and F1-score are all 0.9841, 0.9837, 0.9898, and 0.9867, respectively.

Output class	pneumonia	376 60.26%	6 0.96%	98.43% 1.57%
	normal	4 0.64%	238 38.14%	98.35% 1.65%
		98.95% 1.05%	97.54% 2.46%	98.4% 1.6%
	pneumonia	normal		
			Target class	

Fig. 5. Confusion matrix.

The Receiver Operating Characteristic (ROC) curve for the suggested model is depicted in Fig. 6. The curve illustrates a critical performance indicator for every classification model. The AUC is approximately one (0.991 for pneumonia and 0.986 for normal) and the slope in the graph is extremely close to the upper left corner, showing excellent result in differentiating between the two categories. The curve also demonstrates that the suggested model's capacity to distinguish between pediatric pneumonia and normal is nearly comparable.

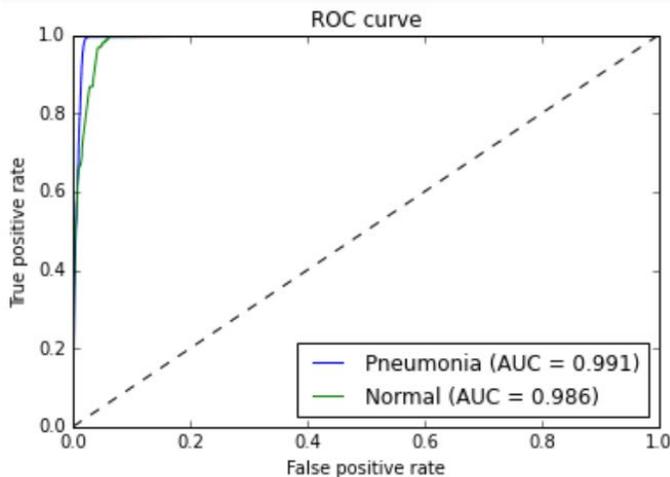


Fig. 6. Receiver operating characteristic curve.

### C. Comparative Study with State-of-the-Art Methods

In Table I, a comparison research is shown. The suggested approach is compared against other prevailing techniques' accuracy. In order to classify cases of pneumonia, Liang et al. [15], Zubair et al. [30], and Mahmud et al. [31] employed simple transfer learning techniques and reached accuracy levels of 90.5%, 96.6%, and 98.1%, respectively. Using specially designed CNNs, Rajaraman et al. [12] have been able to attain an accuracy of 96.2%. An accuracy of 96.36% has been obtained by Chouhan et al. [32] who developed an ensemble model to aggregate estimates from various deep learning

algorithms. The quantity of deep features has been decreased by Togacar et al. [19], who claimed accuracy of 96.84%. An accuracy of 98.0% was attained by Rahman et al. [18] using CNNs. Hashmi et al. [33] attained an accuracy of 98.3%. The strategy proposed in current research performed better than the existing approaches and achieved an accuracy of 98.4% with precision, recall, and AUC of 0.9898, 0.9841, and 0.991, respectively.

TABLE I. COMPARISON WITH STATE-OF-THE-ART METHODS

Model	Accuracy	Precision	Recall	AUC
Liang et al. [15]	90.50%	0.891	0.967	0.927
Rajaraman et al. [12]	96.20%	0.962	0.995	0.993
Chouhan et al. [32]	96.39%	0.9328	0.996	0.9934
Zubair et al. [30]	96.60%	0.972	0.981	-
Togacar et al. [19]	96.84%	0.9688	0.9683	0.968
Rahman et al. [18]	98.00%	0.97	0.99	0.98
Mahmud et al. [31]	98.10%	0.98	0.985	0.983
Hashmi et al. [33]	98.30%	0.9825	0.9814	0.9971
Proposed Model	98.40%	0.9898	0.9841	0.991

### V. CONCLUSIONS

The most common cause of mortality for kids under the age of five globally is pneumonia. Chest X-rays are evaluated by qualified radiologists to diagnose pneumonia. However, it takes a lot of time and is tiresome. Methodologies for biomedical image diagnostics have a lot of potential for use in clinical image analysis. In this study, authors provide an efficient deep learning algorithm for more accurate pediatric pneumonia detection from CXR scans. By using sophisticated pre-processing techniques to render CXR scans more visible, the suggested method collects low-level feature space, enabling recognition of complicated trends from clinical scans on a par with skilled radiologists. The suggested approach may aid in illness detection and support radiologists in the patient care. To evaluate the effectiveness of the suggested model, several metrics, including accuracy, sensitivity, specificity, precision, and AUC score, are calculated. On the testing dataset, the suggested model achieved a precision of 0.9898, an F1-score of 0.9867, an AUC score of 0.991, a sensitivity of 0.9841, a specificity of 0.9837, and an accuracy of 98.4%. Furthermore, based on the current evaluation criteria, the suggested model beats cutting-edge architectures.

### REFERENCES

- [1] I. Rudan, C. Boschi-Pinto, Z. Biloglav, K. Mulholland and H. Campbell, "Epidemiology and etiology of childhood pneumonia," *Bulletin of the world health organization*, vol. 86, pp. 408-416B, 2008.
- [2] T. Cherian, E. K. Mulholland, J. B. Carlin, H. Ostensen, R. Amin, M. D. Campo, D. Greenberg, R. Lagos, M. Lucero, S. A. Madhi and K. L. O'Brien, "Standardized interpretation of paediatric chest radiographs for the diagnosis of pneumonia in epidemiological studies," *Bulletin of the World Health Organization*, vol. 83, pp. 353-359, 2005.
- [3] N. Liu, L. Wan, Y. Zhang, T. Zhou, H. Huo and T. Fang, "Exploiting convolutional neural networks with deeply local description for remote sensing image classification," *IEEE access*, vol. 6, pp. 11215-11228, 2018.

- [4] S. Batra, R. Khurana, M. Z. Khan, W. Boulila, A. Koubaa and P. Srivastava, "A Pragmatic Ensemble Strategy for Missing Values Imputation in Health Records," *Entropy*, vol. 24, no. 4, p. 533, 2024.
- [5] J. W. Picone, "Signal modeling techniques in speech recognition," *Proceedings of the IEEE*, vol. 81, no. 9, pp. 1215-1247, 1993.
- [6] S. Hazra and A. Santra, "Robust gesture recognition using millimetric-wave radar system," *IEEE sensors letters*, vol. 2, no. 4, pp. 1-4, 2018.
- [7] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*, 2009.
- [8] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghighi, R. Ball, K. Shpanskaya and J. Seekins, "Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison," in *Proceedings of the AAAI conference on artificial intelligence*, 2019.
- [9] P. Rajpurkar, J. Irvin, R. L. Ball, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. P. Langlotz and B. N. Patel, "Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists," *PLoS medicine*, vol. 15, no. 11, p. e1002686, 2018.
- [10] S. Batra and S. Sachdeva, "Organizing standardized electronic healthcare records data for mining," *Health Policy and Technology*, vol. 5, no. 3, pp. 226-242, 2016.
- [11] D. S. Kermany, M. Goldbaum, W. Cai, C. C. Valentim, H. Liang, S. Baxter, A. McKeown, G. Yang, X. Wu, F. Yan and J. Dong, "Identifying medical diagnoses and treatable diseases by image-based deep learning," *Cell*, vol. 172, no. 5, pp. 1122-1131, 2018.
- [12] S. Rajaraman, S. Candemir, I. Kim, G. Thoma and S. Antani, "Visualization and interpretation of convolutional neural network predictions in detecting pneumonia in pediatric chest radiographs," *Applied Sciences*, vol. 8, no. 10, p. 1715, 2018.
- [13] P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. Langlotz, K. Shpanskaya and M. P. Lungren, "Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning," *arXiv preprint*, 2017.
- [14] O. Stephen, M. Sain, U. J. Maduh and D. U. Jeong, "An efficient deep learning approach to pneumonia classification in healthcare," *Journal of healthcare engineering*, 2019.
- [15] G. Liang and L. Zheng, "A transfer learning method with deep residual network for pediatric pneumonia diagnosis," *Computer methods and programs in biomedicine*, vol. 187, p. 104964, 2020.
- [16] R. Siddiqi, "Automated pneumonia diagnosis using a customized sequential convolutional neural network," in *Proceedings of the 2019 3rd international conference on deep learning technologies*, 2019.
- [17] X. Gu, L. Pan, H. Liang and R. Yang, "Classification of bacterial and viral childhood pneumonia using deep learning in chest radiography," in *Proceedings of the 3rd international conference on multimedia and image processing*, 2018.
- [18] T. Rahman, M. E. Chowdhury, A. Khandakar, K. R. Islam, K. Islam, Z. B. Mahbub, M. A. Kadir and S. Kashem, "Transfer learning with deep convolutional neural network (CNN) for pneumonia detection using chest X-ray," *Applied Sciences*, vol. 10, no. 9, p. 3233, 2020.
- [19] M. Toğaçar, B. Ergen, Z. Cömert and F. Özyurt, "A deep feature learning model for pneumonia detection applying a combination of mRMR feature selection and machine learning models," *Irbm*, vol. 41, no. 4, pp. 212-222, 2020.
- [20] A. Mittal, D. Kumar, M. Mittal, T. Saba, I. Abunadi, A. Rehman and S. Roy, "Detecting pneumonia using convolutions and dynamic capsule routing for chest X-ray images," *Sensors*, vol. 20, no. 4, p. 1068, 2020.
- [21] "White Balance," [Online]. Available: <https://docs.gimp.org/2.10/en/gimp-layer-white-balance.html>. [Accessed 18 November 2022].
- [22] A. M. Reza, "Realization of the contrast limited adaptive histogram equalization (CLAHE) for real-time image enhancement," *Journal of VLSI signal processing systems for signal, image and video technology*, vol. 38, pp. 35-44, 2004.
- [23] S. M. Pizer, E. P. Amburn, J. D. Austin, R. Cromartie, A. Geselowitz, T. Greer, B. ter Haar Romeny, J. B. Zimmerman and K. Zuiderveld, "Adaptive histogram equalization and its variations," *Computer vision, graphics, and image processing*, vol. 39, no. 3, pp. 355-368, 1987.
- [24] S. M. Pizer, "Contrast-limited adaptive histogram equalization: Speed and effectiveness stephen m. pizer, r. eugene johnston, james p. ericksen, bonnie c. yankaskas, keith e. muller medical image display research group," in *Proceedings of the first conference on visualization in biomedical computing*, Atlanta, Georgia, 1990.
- [25] S. Batra, H. Sharma, W. Boulila, V. Arya, P. Srivastava, M. Z. Khan and M. Krichen, "An Intelligent Sensor Based Decision Support System for Diagnosing Pulmonary Ailment through Standardized Chest X-ray Scans," *Sensors*, vol. 22, no. 19, p. 7474, 2022.
- [26] A. Pathak, S. Batra and V. Sharma, "An Assessment of the Missing Data Imputation Techniques for COVID-19 Data," in *Proceedings of 3rd International Conference on Machine Learning, Advances in Computing, Renewable Energy and Communication: MARC 2021*, 2022.
- [27] J. Wang and L. Perez, "The effectiveness of data augmentation in image classification using deep learning," *Convolutional Neural Networks Vis. Recognit*, vol. 11, no. 2017, pp. 1-8, 2017.
- [28] D. Kermany, K. Zhang and M. Goldbaum, "Labeled optical coherence tomography (oct) and chest x-ray images for classification," *Mendeley data*, vol. 2, no. 2, p. 651, 2018.
- [29] R. C. Hidayatullah and S. Violina, "Convolutional neural network architecture and data augmentation for pneumonia classification from chest X-rays images," *Int J Innov Sci Res Technol*, vol. 5, pp. 158-164, 2020.
- [30] S. H. A. H. Zubair, "An efficient method to predict pneumonia from chest X-rays using deep learning approach," *The Importance Of Health Informatics In Public Health During A Pandemic*, vol. 272, p. 457, 2020.
- [31] T. Mahmud, M. A. Rahman and S. A. Fattah, "CovXNet: A multi-dilation convolutional neural network for automatic COVID-19 and other pneumonia detection from chest X-ray images with transferable multi-receptive feature optimization," *Computers in biology and medicine*, vol. 122, p. 103869, 2020.
- [32] V. Chouhan, S. Singh, A. Khamparia, D. Gupta, P. Tiwari, C. Moreira, R. Damaševičius and V. H. C. De Albuquerque, "A novel transfer learning based approach for pneumonia detection in chest X-ray images," *Applied Sciences*, vol. 10, no. 2, p. 559, 2020.
- [33] M. F. Hashmi, S. Katiyar, A. W. Hashmi and A. G. Keskar, "Pneumonia detection in chest X-ray images using compound scaled deep learning model," *Automatika*, vol. 62, no. 3-4, pp. 397-406, 2021.

# Long Short-Term Memory for Non-Factoid Answer Selection in Indonesian Question Answering System for Health Information

Retno Kusumaningrum\*, Alfi F. Hanifah, Khadijah Khadijah, Sukmawati N. Endah, Priyo S. Sasongko  
Department of Informatics, Universitas Diponegoro, Semarang, Indonesia

**Abstract**—Providing reliable health information to a community can help raise awareness of the dangers of diseases, their causes, methods of prevention, and treatment. Indonesians are facing various health problems partly due to the lack of health information; hence, the community needs media that can effectively provide reliable health information, namely a question answering (QA) system. The frequently asked questions are non-factoid questions. The development of answer selection based on the classical approach requires distinctive engineering features, linguistic tools, or external resources. It can be solved using deep learning approach such as Convolutional Neural Networks (CNN). However, this model cannot capture the sequence of words in both questions and answers. Therefore, this study aims to implement a long short-term memory (LSTM) model to effectively exploit long-range sequential context information for an answer selection task. In addition, this study analyses various hyper-parameters of Word2Vec and LSTM, such as the dimension, context window, dropout, hidden unit, learning rate, and margin; the corresponding values that yield the best mean reciprocal rank (MRR) and mean average precision (MAP) are found to be 300, 15, 0.25, 100, 0.01, and 0.1, respectively. The best model yields MAP and MRR values of 82.05% and 91.58%, respectively. These results experienced an increase in MAP and MRR of 18.68% and 46.11%, respectively, compared to CNN as the baseline model.

**Keywords**—Answer selection; health information; long short-term memory; LSTM; question answering

## I. INTRODUCTION

Providing reliable health information to a community can help raise awareness of the dangers of diseases, their causes, methods of prevention, and treatment. Indonesians are facing various health problems partly due to the lack of health information, including the dangers of smoking, nutritional problems (stunting and obesity), and serious diseases such as heart disease, cancer, and diabetes. Therefore, the community requires media that can provide health information appropriately, namely a question answering (QA) system.

The QA system is a natural language processing (NLP) application that provides specific answers to the questions/queries posed by the user. The QA system is different from a search engine in that the latter will return a set of documents that may contain answers, and users are required to read the documents and search for the exact answers or infer from the set of documents presented. Therefore, the process of finding answers in a QA system is more complex

than the process of finding documents presented by a search engine.

Various QA systems have been developed for both the non-Indonesian QA system and the Indonesian QA system. The following QA systems have been developed for non-Indonesian documents: the English QA system [1]–[3], Chinese QA system [4], Spanish QA system [5]–[7], and French QA system [8], [9]. The Indonesian QA system includes QA statistical and linguistic knowledge systems [10], syntactic-semantic processing QA systems [11], [12], QA systems based on machine learning cross-language QA systems [13], pattern matching QA-based systems [14], and pipeline-based cross-language QA systems [15]. In addition, the Indonesian language QA system has been developed for closed-domain QA [16]–[18].

QA systems are differentiated on the basis of the type of questions handled, which are divided into five categories: factoid, non-factoid, yes-no, list, and opinion [19]. Factoid questions have answers in the form of date, quantity, location, person, organisation, and name (in the form of nouns) in addition to the location, person, and organisation categories [13]. Non-factoid questions are those whose answers are generally used to understand something. Non-factoid questions have six categories: question definitions, reasons, methods, degrees, changes, and details [20]. Overall, the Indonesian QA system is still limited to factoid questions, with hardly any non-factoid questions. Related to health information, the types of questions that are commonly encountered are non-factoid questions.

Several studies have been conducted on non-factoid Indonesian QA systems but for non-health data domains. Moreover, these studies generally used a classical approach such as pattern matching and semantic analysis [21], case-based reasoning [16], and similarity score technique [19]. They provide a good performance only when all the patterns of the answer pairs have been defined, making it appropriate only for certain knowledge domains. In addition, the studies were generally implemented for non-factoid questions related to definitions, reasons, and method categories.

Now-a-days, deep learning models have been widely developed for solving several problems using various types of datasets, such as those containing images, signals, and text. Some examples of deep learning implementation using textual data include sentiment analysis [22], [23], machine translation [24], [25], summarisation [26], [27], and QA. A deep learning

model can be implemented in a QA system as a model for selecting the exact answer from a set of candidate answers, also known as the answer pool. The deep learning model does not require feature engineering, linguistic tools, or external resources [28]. Feature engineering is the stage wherein representative features, such as term frequency-inverse document frequency (TF-IDF) and bag-of-words, are determined. The linguistic tools are linguistic rules and syntax. The implementation of deep learning in a QA system requires a convolutional neural network (CNN). However, this model cannot capture the sequence of words in both questions and answers. This can be overcome by implementing long short-term memory (LSTM).

Therefore, this study aims to implement an LSTM as a model for selecting non-factoid answers in the Indonesian question answering system (IQAS) for Health Information. As mentioned earlier, the LSTM model has never been implemented for answer selection in the IQAS, neither in a specific data domain nor in the general data domain. Hence, the first step in this approach is to train the word2vec model on a health information corpus obtained from various popular health websites written in the Indonesian language. In addition, this study empirically analyses the effect of Word2Vec hyper-parameters, such the dimensions and the context window size, on the performance of the LSTM model in selecting the right answer to a question. Furthermore, the effect of varying the LSTM hyper-parameters on the performance of the LSTM model as a model for selecting exact answers from an answer candidate pool was studied; thus, we established the best answer selection model.

The contributions of this paper are summarised as follows:

- A pre-trained Word2Vec model for the Indonesian language, specifically on health information.
- An investigation related to the influences of the dimensions and context window size of Word2Vec on the performance of the LSTM model in selecting answers.
- An analysis of the influences of the hyper-parameters on the LSTM model, including the dropout, number of hidden units, learning rate, and margin size, on the performance of the LSTM model for answer selection.
- A pre-trained LSTM model for non-factoid answer selection in the IQAS for health information. Subsequently, it was implemented as a web-based application.

The rest of this paper is organised as follows. Section II describes related work, including a general description of the answer selection task and LSTM in detail. A detailed explanation of the proposed framework is presented in Section III, including descriptions of data collection, training process of Word2Vec, generation of the answer selection model based on the LSTM, and model evaluation. Section IV presents the experimental results. Finally, in Section V, we draw some conclusions from the results.

## II. RELATED WORKS

### A. Answer Selection Task

Answer selection is a subtask of the QA system that performs the process of selecting sentences containing the required information from a set of candidate answers [29]. Answer selection involves not only matching the terms in the question and answer but also finding the same semantic meaning from both the question and answer. Formally, the answer selection problem can be described as follows:

- There is a question  $q$  and answer candidate pool  $\{a_1, a_2, \dots, a_s\}$  that contains a set of answer candidates for a particular question.
- The aim of answer selection is to select the best answer candidates from the answer candidate pool.

Therefore, the answer selection task can be formulated as a ranking problem, giving better ranks to answers that are more relevant to the respective question. Some of the ranking function approaches include pointwise, pairwise, and list wise [30]. This study implements a pairwise approach to train the ranking function to give higher scores for correct answers and lower scores for wrong ones.

### B. Long Short-Term Memory

The LSTM model is a popular variation of the recurrent neural network (RNN) method. The RNN method is widely used to solve data problems whose order requires attention. The LSTM model overcomes the gradient vanishing problem of the RNN method. In addition, LSTM model is more capable of dealing with the context of long and sequential information. The LSTM model used in this study is the one introduced in [31].

The LSTM model is designed to solve the gradient vanishing problem using a gate mechanism. Its architecture has three gates, namely an input gate  $it$ , a forget gate  $ft$ , and an output gate  $ot$ , and a memory cell  $Ct$ . The LSTM can add or reduce information into the cell state, which is regulated by the gate. The input gate is responsible for determining new information to be added to the memory cell. The forget gate determines which information will be saved or deleted. Finally, the output gate is responsible for determining the information that will be used as output. Fig. 1 shows the LSTM cells.

The hidden state  $ht$  is calculated on the basis of the three LSTM gates. The size of the hidden state is determined by a parameter called the hidden unit. The hidden unit is a parameter in the LSTM that shows the vector dimension of the hidden state  $ht$  for each time step. Mathematically, the LSTM model is defined as follows:

$$i_t = \sigma(W_i x(t) + U_i h(t-1) + b_i) \quad (1)$$

$$f_t = \sigma(W_f x(t) + U_f h(t-1) + b_f) \quad (2)$$

$$o_t = \sigma(W_o x(t) + U_o h(t-1) + b_o) \quad (3)$$

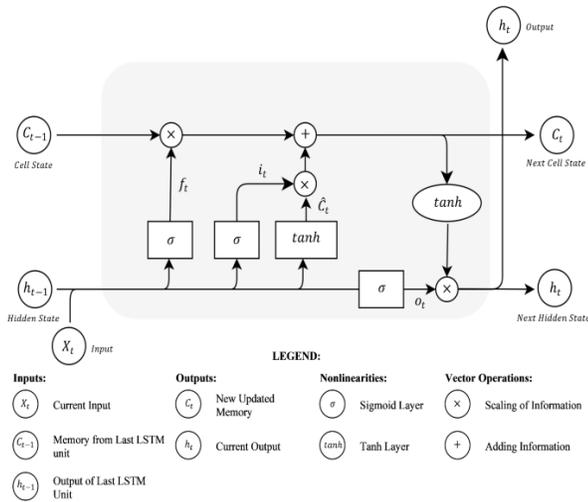


Fig. 1. LSTM cell

$$\tilde{c}_t = \tanh(W_c x(t) + U_c h(t-1) + b_c) \quad (4)$$

$$c_t = i_t * \tilde{c}_t + f_t * c_{t-1} \quad (5)$$

$$h_t = o_t * \tanh(c_t) \quad (6)$$

The LSTM architecture has three gates (input  $i$ , forget  $f$ , and output  $o$ ) and a cell memory vector  $c$ .  $\sigma$  is the sigmoid function.  $W$ ,  $U$ , and  $b$  are the network parameters.

### III. METHODOLOGY

This section describes the proposed framework used in this study, comprising four main processes. Fig. 2 shows its general description.

The research framework comprises four main processes: data collection, training process of Word2Vec, generating an answer selection model based on the LSTM, and model evaluation. The detailed explanations for each process are given in the following subsections.

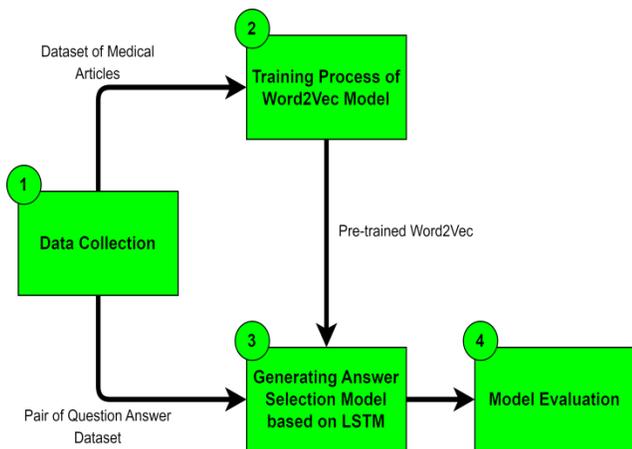


Fig. 2. Framework of this study comprising four main processes: data collection, word2vec training, LSTM-based answer selection modelling, and model evaluation

### A. Data Collection

In this process, two types of datasets are formed: a QA dataset (pair of question-and-answer datasets) and a health article dataset. The QA dataset was created by collecting question and answer pairs from popular health sites in Indonesia, namely hellosehat.com, alodokter.com, and halodoc.com. Non-factoid questions on topics of diseases and medicines are used as questions. The categories of the questions are definitions, reasons, and methods. In total, 750 pairs of questions and answers are formed, consisting of 355 pairs for definitions, 145 pairs for reasons, and 250 pairs for methods. The article dataset is established using all the articles from the three websites through data scraping.

### B. Training Process of Word2Vec Model

The Word2Vec model is a word embedding algorithm proposed in [32] to learn vector representations. Vector representations can efficiently capture the semantic meaning of the words represented. The word vector tends to obey the laws of analogy and describe intuition. Words known as synonyms have the same vector in the cosine equation, whereas antonyms have different vectors. Therefore, the representation of words in the vector space is useful for achieving better performance on NLP problems by grouping similar words.

The dataset used in Word2Vec training is the article dataset. The article dataset contains articles on diseases and medicines found on the three sites previously described. The number of vocabularies formed was 44,700. The Word2Vec model used is skip-gram, and the evaluation method is hierarchical softmax. Fig. 3 illustrates the skip-gram architecture.

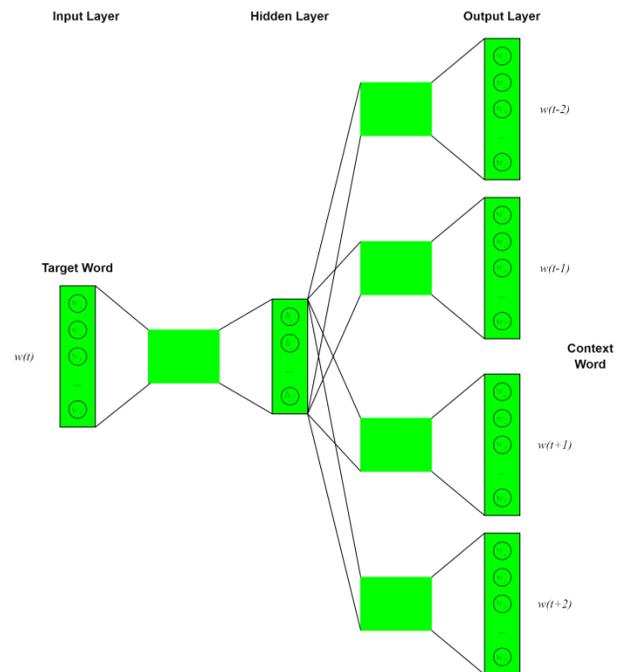


Fig. 3. Illustration of skip-gram architecture of Word2Vec model

### C. Generating Answer Selection Model based on LSTM

Modelling for answer selection uses a Siamese architecture. This type of architecture can be used to measure the relevance of candidate answers to a question. Fig. 4 shows the Siamese architecture of the LSTM-based answer selection model. In the embedding layer, the inputted sentences (i.e., the candidate answer and the question) are converted into vector representations generated by Word2Vec training. Thereafter, in the encoding layer, the same encoder is used to create distributed vector representations for the input sentences separately. The encoding layer adopts the QA-LSTM using a bidirectional LSTM (biLSTM) model. During the encoding process, the questions and answers do not have explicit interactions.

Bidirectional LSTM utilises both the previous and future contexts by processing in two directions and generates two independent sequences of LSTM output vectors. The two output vectors are concatenated as follows:

The implementation of max pooling was used to generate representations for the questions and answers based on the word-level biLSTM outputs. The relevance scores of the candidate answers to a question are obtained based on pooled vectors. Subsequently, using the cosine similarity measures the distance between the candidate's answer and the question.

### D. Model Evaluation

The evaluation techniques used are the mean reciprocal rank (MRR) and mean average precision (MAP), which are the standard metrics for information retrieval and QA. The MRR can be calculated as follows:

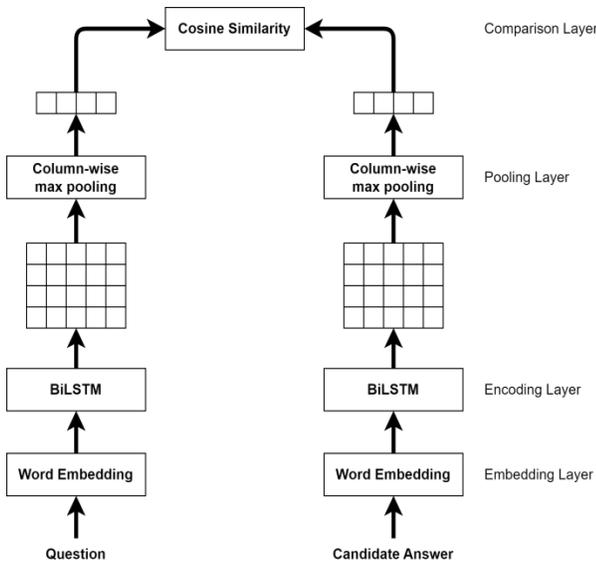


Fig. 4. Siamese architecture of LSTM-based answer selection

$$MRR = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{rank_i} \quad (7)$$

The MAP can be calculated as follows:

$$MAP = \frac{1}{|Q|} \sum_{j=1}^{|Q|} \frac{1}{m_j} \sum_{k=1}^{|m_j|} Precision(R_{jk}) \quad (8)$$

## IV. RESULTS AND DISCUSSION

### A. Experimental Setup

The data used in this research are in the form of 750 question-answer pairs. There are 1564 unique answers collected in the answer space. With regard to the distribution ratio of the training and test data, 70% is for training and 30% is for testing. Following the data distribution, we have 525 pairs as training data and 225 pairs as test data. The pool size is 50. It was generated by sending the ground-truth answers to the pool and randomly sampling negative answers from the answer space until the pool size reached 50.

The experiment employs several hyperparameters of the Word2Vec model and LSTM. Each model is trained for 100 epochs. The Word2Vec hyperparameters are dimension (100, 200, and 300) and context window (5, 10, and 15). At the same time, the LSTM hyperparameters are dropout (0.25, 0.5, and 0.75), number of hidden units (50, 75, and 100), learning rate (0.00001, 0.0001, 0.001, and 0.01), and margin (0.05, 0.1, and 0.15).

### B. Experimental Scenarios

Several scenarios are established to determine the impacts of the various parameters tested on the performance of the proposed model; scenarios 1, 2, 3, 4, 5, and 6 are for the Word2Vec dimension, context window, dropout, hidden unit, learning rate, and margin, respectively. Fig. 5 shows the overview of these scenarios.

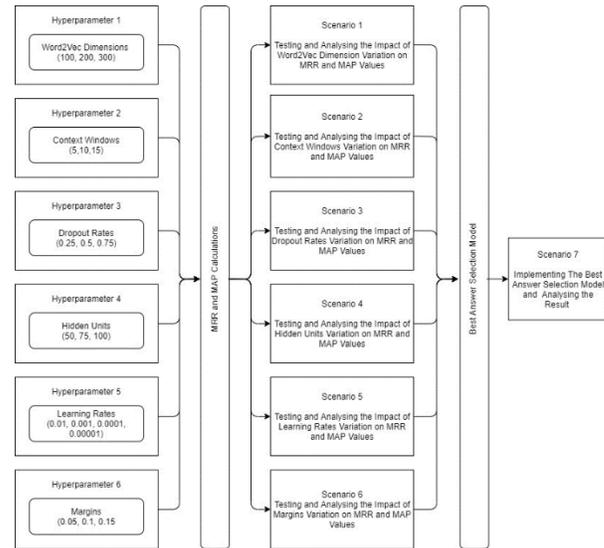


Fig. 5. Six hyperparameters are tested. Each combination produces a model that calculates the MRR and MAP values. The overall results are analysed through seven scenarios. The best model is obtained from the model that produces the best MRR and MAP values

### C. Experimental Results and Analysis

Scenario 1 is aimed at studying the impact of Word2Vec dimensions on the MRR and MAP results. Table I shows that the model yields the best averages of MRR (78.75%) and MAP (63.70%) when the Word2Vec dimension is 300. The MRR and MAP values are directly proportional to the dimensions of Word2Vec; therefore, the higher the dimensions of Word2Vec, the higher the MRR and MAP

values. The Word2Vec dimension represents the size of the learned word vector, or it can be referred to as the features of each word. A higher dimension tends to capture more information and better word representations.

Scenario 2 is aimed at studying the impacts of context window on the MRR and MAP results. The best averages of the MRR and MAP values are obtained when the context window is 15, as shown in Table II. From the table, it can be concluded that the averages of the MRR and MAP are directly proportional to the context window, which means that, the larger the context window size, the higher the average MRR and MAP values. The size of the context window defines the range of words to be included as the context of a target word. For instance, a window size of 5 takes five words before and after a target word as its context for training. A larger context window is required to answer non-factoid questions on health information because this type of question requires a longer answer. Moreover, answers related to health information typically have a long explanation.

Scenario 3 is aimed at studying the impacts of dropout rate on the MRR and MAP results. The best MRR and MAP values are 81.25% and 66.58% when the dropout value is set to 0.25. From the average MRR and MAP obtained for all the tested dropout values, it can be concluded that the dropout value is inversely proportional to the average MRR and MAP, which means that, the lower the dropout value, the higher the MRR and MAP. Dropout refers to ignoring units (i.e. neurons) during the training phase of a certain set of neurons. A higher dropout value indicates that more neurons are ignored, and this will cause the model to lose its ability to learn. Moreover, the dropout performed on the LSTM model can make the model to be more limited in keeping the memory. Therefore, lower dropouts are considered better for storing memory in the LSTM model. Table III lists the results of scenario 3.

TABLE I. PERFORMANCE COMPARISON WHEN VARYING THE WORD2VEC DIMENSIONS

Dimension	Average of MAP (%)	Average of MRR (%)
100	56.76	72.91
200	61.77	77.23
300	63.70	78.75

TABLE II. PERFORMANCE COMPARISON WHEN VARYING THE CONTEXT WINDOWS

Context Window	Average of MAP (%)	Average of MRR (%)
5	58.24	74.30
10	61.15	76.63
15	62.84	77.96

TABLE III. PERFORMANCE COMPARISON WHEN VARYING THE DROPOUT RATES

Dropout Rate	Average of MAP (%)	Average of MRR (%)
0.25	66.58	81.25
0.5	62.14	77.50
0.75	53.51	70.14

Scenario 4 is aimed to study the impacts of hidden units on the MRR and MAP results. As mentioned before, this study applies different numbers of hidden units: 50, 75, and 100. From Table IV, it can be concluded that the number of hidden units is directly proportional to the average MRR and MAP. The output dimension determines the number of dimensions for each word in the input sequence. Dimension implies the number of features to be remembered. The best averages of MRR and MAP are obtained under a hidden unit value of 100. This is because using more features provides a better representation than using fewer features.

Scenario 5 is aimed at studying the impacts of the learning rate on the MRR and MAP results. Several learning rates were set: 0.01, 0.001, 0.0001, and 0.00001. Based on Table V, it can be concluded that the learning rate is directly proportional to the averages of MRR and MAP. The best averages of MRR and MAP are obtained under a learning rate of 0.01. As explained in the experimental results section, all the models are trained for 100 epochs. The learning rate is a hyperparameter that helps control the degree of model change. A low learning rate may result in a long training process that could get stuck, making it difficult to converge. These results can be obtained because the epoch used tends to be small; therefore, a high learning rate will decrease the MRR and MAP values.

Scenario 6 is aimed at studying the impact of margin on the MRR and MAP results. As previously explained, there are three different margin values: 0.05, 0.1, and 0.15. The highest average MRR and MAP were obtained under a margin of 0.1, as listed in Table VI. No specific pattern is generated between the margins with the average MRR and MAP. Margin is a variable in the hinge loss function. The hinge loss function is an employed loss function that was minimised in this research. If the ground-truth answer has a score higher than the negative answer by at least a margin, the expression has a zero loss. Condition here implies margins as the optimum distance that can be produced between the ground-truth answer and negative answers. If the margin value is too low, the ground-truth answer and the negative answer will not be separated appropriately. The lower the margin, the smaller the distance between the ground-truth and negative answers. This condition can make relevant answers irretrievable. Meanwhile, if the margin is too high, the distance between the correct answer and the wrong answer will be even greater. This makes irrelevant answers be incorrectly taken as correct answers.

TABLE IV. PERFORMANCE COMPARISON WHEN VARYING THE HIDDEN UNITS

Hidden Units	Average of MAP (%)	Average of MRR (%)
50	58.16	74.15
75	61.12	76.62
100	62.95	78.11

TABLE V. PERFORMANCE COMPARISON WHEN VARYING THE LEARNING RATES

Learning Rates	Average of MAP (%)	Average of MRR (%)
0.00001	50.74	68.41
0.0001	54.14	70.95
0.001	62.56	78.10
0.01	75.53	87.72

TABLE VI. PERFORMANCE COMPARISON WHEN VARYING THE MARGINS

Margins	Average of MAP (%)	Average of MRR (%)
0.05	60.80	76.39
0.1	61.02	76.55
0.15	60.41	75.94

Based on the results of scenarios 1 to 6, the best answer selection model is obtained when using the following hyperparameters: word2vec dimension is 300, context window size is 15, dropout rate value is 0.25, number of hidden units is 100, learning rate is 0.01, and margin value is 0.1. This model yields MAP and MRR values of 82.05% and 91.58%, respectively.

Compared with previous research, this study also run experiments using CNN with an architecture consisting of 4 convolution layers (kernel size in 1, 2, 3, and 5) and one pooling layer. The word2vec dimension used in the test uses the same dimension, namely 300. The best parameter results for the CNN model include margin 0.15, hidden unit 100, dropout 0.25, learning rate 0.01, and context window 15. The MAP and MRR values obtained are 63.37% and 45.47%, respectively. An illustration of the difference between the CNN model and the proposed model can be seen in Fig. 6. It can be seen that the increases in MAP and MRR were 18.68% and 46.11%, respectively.

Subsequently, the best model is implemented for the QA application, which is given the name MediQA. Fig. 7 shows the sample result of the answer selection.

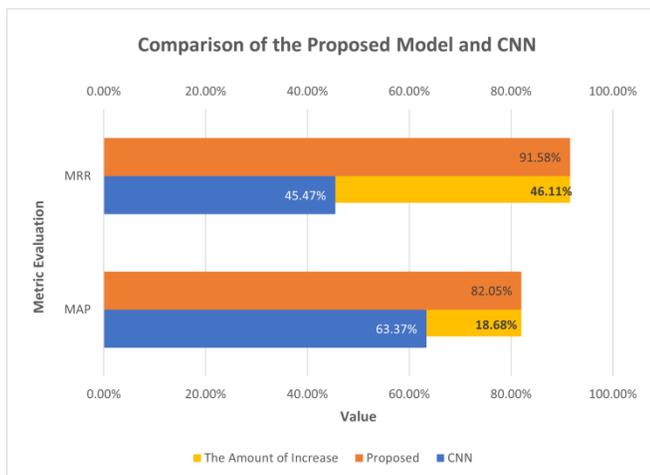


Fig. 6. Comparison of the proposed model and CNN (as baseline model)

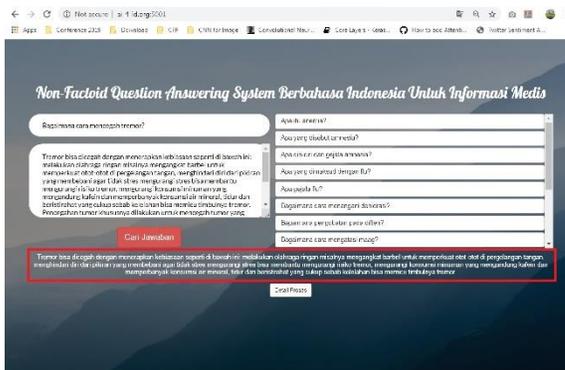


Fig. 7. Siamese architecture of LSTM-based answer selection

**Incorrect Answer**

**Question:**  
Apa itu tetanus? (*What is tetanus?*)

**Answer Pool:**  
Tetanus adalah kondisi kaku dan tegang di seluruh tubuh akibat infeksi kuman. Kaku dan tegang seluruh tubuh ini terasa menyakitkan dan dapat menyebabkan kematian. **Gejala tetanus akan muncul dalam 4-21 hari setelah terinfeksi.** Kuman atau bakteri tetanus masuk ke dalam tubuh melalui luka pada kulit, dan akan mengeluarkan racun untuk menyerang saraf. Bakteri ini bernama Clostridium tetani, yang banyak ditemukan pada tanah, debu, atau kotoran hewan. (*Tetanus is a stiff and tense condition throughout the body due to germ infection. Stiffness and tension throughout the body are painful and can cause death. Symptoms of tetanus will appear within 4-21 days after infection. Tetanus bacteria or bacteria enter the body through wounds on the skin and will release toxins to attack the nerves. This bacterium is called Clostridium tetani, which is found mostly in soil, dust, or animal feces.*)

**Selected Answer:**  
Gejala tetanus akan muncul dalam 4-21 hari setelah terinfeksi. (*Symptoms of tetanus will appear within 4-21 days after infection.*)

**Correct Answer:**  
Tetanus adalah kondisi kaku dan tegang di seluruh tubuh akibat infeksi kuman. (*Tetanus is a stiff and tense condition throughout the body due to germ infection.*)

(a)

---

**Correct Answer**

**Question:**  
Apa obat eksim? (*What is eczema medicine?*)

**Answer Pool:**  
Dokter biasanya akan melakukan diagnosis dengan memeriksa kulit dan sejarah medis anda. Dokter juga dapat menggunakan tes patch atau tes lainnya untuk mengeliminasi penyakit kulit lainnya atau mengidentifikasi kondisi yang menyertai eksim. **Obat-obatan yang umum untuk perawatan eksim meliputi hydrocortisone, antihistamin, corticosteroid.** (*The doctor will usually make a diagnosis by examining your skin and medical history. Doctors can also use patch tests or other tests to eliminate other skin diseases or identify the conditions that accompany eczema. Common medications for the treatment of eczema include hydrocortisone, antihistamines, corticosteroids.*)

**Selected Answer:**  
Obat-obatan yang umum untuk perawatan eksim meliputi hydrocortisone, antihistamin, corticosteroid. (*Common medications for the treatment of eczema include hydrocortisone, antihistamines, corticosteroids.*)

(b)

Fig. 8. Sample result of answer selection of definition question, (a) Sample of incorrect answer, (b) Sample of correct answer

As mentioned in the previous section, this study evaluates three questions: definitions, reasons, and methods. Fig. 8 shows a sample of the correct and incorrect answer results given by the MediQA application for the definition question type. Fig. 9 shows the same for the method question type. Both figures consist of two parts, the first part shows a result example of choosing the incorrect answer by the system, and the second part shows a result example of choosing the correct answer by the system. In the answer pool section, sentences in green indicate sentences that should have been selected as the correct answer. Meanwhile, sentences written in red are incorrect answer sentences and are output as answers by the system.

Incorrect Answer
<b>Question:</b> Bagaimana cara mengonsumsi disulfiram? (How to take disulfiram?)
<b>Answer Pool:</b> Dosis akan disesuaikan berdasarkan kondisi pasien, dengan jangka waktu pengobatan kurang dari 6 bulan. Obat ini dapat dikonsumsi sebelum atau setelah makan di pagi hari. <b>Berikut adalah beberapa efek samping yang dapat terjadi setelah menggunakan disulfiram: pusing, mudah merasa lelah, muncul jerawat, mulut terasa tidak enak (seperti rasa bawang atau metal).</b> (The dosage will be adjusted based on the patient's condition, with a treatment period of fewer than 6 months. This medicine can be taken before or after eating in the morning. <b>Here are some of the side effects that can occur after using disulfiram: dizziness, easy feeling tired, pimples appear, the mouth feels bad (like the taste of onions or metal).</b>
<b>Selected Answer:</b> Disulfiram adalah beberapa efek samping yang dapat terjadi setelah menggunakan disulfiram: pusing, mudah merasa lelah, muncul jerawat, mulut terasa tidak enak (seperti rasa bawang atau metal). (Here are some of the side effects that can occur after using disulfiram: dizziness, easy feeling tired, pimples appear, the mouth feels bad (like the taste of onions or metal).)
<b>Correct Answer:</b> Obat ini dapat dikonsumsi sebelum atau setelah makan di pagi hari. (This medicine can be taken before or after eating in the morning)

(a)

Correct Answer
<b>Question:</b> Bagaimana cara mengobati sakit tenggorokan? (How to treat a sore throat?)
<b>Answer Pool:</b> Sakit tenggorokan umumnya dapat diobati dengan banyak minum air dingin dan beristirahat yang cukup. Namun bila gejala masih berlangsung, segera ke dokter agar mendapat penanganan yang tepat. Dokter dapat melakukan pengobatan dengan memberikan obat parasetamol untuk meredakan rasa nyeri di tenggorokan atau memberikan obat antibiotik jika sakit tenggorokan disebabkan oleh infeksi bakteri. Ada beberapa cara untuk mencegah sakit tenggorokan, di antaranya: hindari berbagi perlengkapan makan dan minum dengan orang lain, hindari kontak dengan orang yang sakit, mencuci tangan sebelum makan. (A sore throat can generally be treated by drinking plenty of cold water and getting enough rest. However, if symptoms persist, immediately see the doctor. The doctor can take medication by giving paracetamol to relieve pain in the throat or giving antibiotics if the sore throat is caused by a bacterial infection. There are several ways to prevent sore throats, including: avoid sharing eating and drinking utensils with others, avoid contact with sick people, wash your hands before eating.)
<b>Selected Answer:</b> Dokter dapat melakukan pengobatan dengan memberikan obat parasetamol untuk meredakan rasa nyeri di tenggorokan atau memberikan obat antibiotik jika sakit tenggorokan disebabkan oleh infeksi bakteri. (The doctor can take medication by giving paracetamol to relieve pain in the throat or giving antibiotics if the sore throat is caused by a bacterial infection.)

(b)

Fig. 9. Sample result of answer selection of method question: (a) Sample of incorrect answer, (b) Sample of correct answer

The limitation of this study is that the proposed method focuses on selecting answers on IQAS for a particular domain (health information). At the same time, the need for open-domain QA in Indonesian is still very open. On the other hand, the current state-of-the-art language model reliable for many tasks is Bidirectional Encoder Representations from Transformers (BERT) [33]. The main advantage of BERT is context-sensitive word embedding, where the same word can produce different word embedding when the word has a different context. Word2Vec cannot do this. The Indonesian version of BERT has been developed and is commonly known as IndoBERT [34]. Therefore, it provides an opportunity for further research to apply IndoBERT and LSTM as a model for selecting answers in the Indonesian language open-domain QA.

## V. CONCLUSIONS

This study analyses various hyperparameters of Word2Vec and LSTM applied to non-factoid answer selection in an IQAS for health information. There are six scenarios to evaluate the effects of the hyperparameters on the MRR and MAP

results—first, the larger the dimension of Word2Vec, the better the MRR and MAP values. A dimension of 300 yielded the best MRR and MAP. Second, a context window size of 15 yielded the best MRR and MAP results, indicating that a more extensive context window can yield better MRR and MAP results. Third, a lower dropout value yielded better MRR and MAP values, and the best MRR and MAP were achieved under a dropout value of 0.25. Fourth, the optimum hidden unit value was found to be 100; the higher the number of hidden units, the better the MRR and MAP values. Fifth, a higher learning rate showed significant improvements in the MRR and MAP, given the relatively small number of datasets used in this research. Sixth, a margin of 0.1 produced the best MRR and MAP results. The best model yielded MAP and MRR values of 82.05% and 91.58%, respectively. These results experienced an increase in MAP and MRR of 18.68% and 46.11%, respectively, compared to CNN as the baseline model.

This research is still limited to selecting answers on IQAS for a particular domain (health information), while the need for open-domain QA in Indonesian is still very open. On the other hand, the latest language modelling developments, such as Bidirectional Encoder Representations from Transformers (BERT), have also been developed for Indonesian, commonly known as IndoBERT. Therefore, it provides an opportunity for further research to apply IndoBERT and LSTM as a model for selecting answers in the Indonesian language open-domain QA.

## ACKNOWLEDGMENT

This work was supported in part by the Faculty of Science and Mathematics, Universitas Diponegoro under Grant 4866/UN7.5.8/PP/2019.

## REFERENCES

- [1] M. Kouylekov and B. Magnini, "Recognizing textual entailment with tree edit distance algorithms," in *PASCAL Challenges on RTE*, 2006, pp. 17–20.
- [2] P. Pakray, S. Pal, S. Bandyopadhyay, and A. Gelbukh, "Automatic answer validation system on english language," in *ICACTE 2010 - 2010 3rd International Conference on Advanced Computer Theory and Engineering, Proceedings*, 2010, pp. 329–333.
- [3] S. K. Ray, S. Singh, and B. P. Joshi, "World wide web based question answering system - A relevance feedback framework for automatic answer validation," in *2nd International Conference on the Applications of Digital Information and Web Technologies, ICADIWT 2009*, 2009, pp. 169–174.
- [4] D. Cai, Y. Dong, D. Lv, G. Zhang, and X. Miao, "A web-based Chinese question answering with answering validation," in *Proceedings of 2005 IEEE International Conference on Natural Language Processing and Knowledge Engineering, IEEE NLP-KE'05*, 2005, pp. 499–502.
- [5] Á. Rodrigo, A. Peñas, and F. Verdejo, "The effect of entity recognition in the answer validation," in *CEUR Workshop Proceedings*, 2006, vol. 1172, pp. 1–5.
- [6] A. Téllez-Valero, M. Montes-Y-Gómez, L. Villaseñor-Pineda, and A. Peñas, "Improving question answering by combining multiple systems via answer validation," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 4919 LNCS, pp. 544–554, 2008.
- [7] A. Téllez-Valero, M. Montes-y-Gómez, L. Villaseñor-Pineda, and A. Peñas-Padilla, "Towards multi-stream question answering using answer validation," *Inform.*, vol. 34, no. 1, pp. 45–54, 2010.
- [8] A. L. Ligozat, B. Grau, A. Vilnat, I. Robba, and A. Grappy, "Towards

- an automatic validation of answers in question answering,” in *Proceedings of International Conference on Tools with Artificial Intelligence, ICTAI*, 2007, pp. 444–447.
- [9] A. Grappy, B. Grau, M. H. Falco, A. L. Ligozat, I. Robba, and A. Vilnat, “Selecting answers to questions from Web documents by a robust validation process,” in *Proceedings of 2011 IEEE/WIC/ACM International Conference on Web Intelligence, WI 2011*, 2011, pp. 55–62.
- [10] M. Adriani and S. Adiwibowo, “Finding answers using resources in the internet,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 5152 LNCS, pp. 332–335, 2008.
- [11] S. D. Larasati and R. Manurung, “Towards a semantic analysis of bahasa Indonesia for question answering,” *Proc. 10th Conf. Pacific Assoc. Comput. Linguist.*, pp. 273–280, 2007.
- [12] R. Mahendra, S. D. Larasati, and R. Manurung, “Extending an Indonesian semantic analysis-based question answering system with linguistic and world knowledge axioms,” in *Proceedings of the 22nd Pacific Asia Conference on Language, Information and Computation*, 2008, pp. 262–271.
- [13] A. Purwarianti, M. Tsuchiya, and S. Nakagawa, “A machine learning approach for an Indonesian-English cross language question answering system,” *IEICE Trans. Inf. Syst.*, vol. E90-D, no. 11, pp. 1841–1852, 2007.
- [14] H. Toba and M. Adriani, “Pattern based Indonesian question answering system,” in *Proceedings of the 1st International Conference on Advanced Computer Systems and Information Systems (ICACSIS)*, 2009, pp. 1–6.
- [15] M. I. Faruqi and A. Purwarianti, “An Indonesian question analyzer to enhance the performance of Indonesian-English CLQA,” in *Proceedings of the 2011 International Conference on Electrical Engineering and Informatics, ICEEI 2011*, 2011, pp. K2-1.
- [16] A. Fikri and A. Purwarianti, “Case based Indonesian closed domain question answering system with real world questions,” in *Proceeding of the 7th International Conference on Telecommunication Systems, Services, and Applications, TSSA 2012*, 2012, pp. 181–186.
- [17] A. A. S. Gunawan, P. R. Mulyono, and W. Budiharto, “Indonesian question answering system for solving arithmetic word problems on intelligent humanoid robot,” in *Procedia Computer Science*, 2018, vol. 135, pp. 719–726.
- [18] R. H. Gusmita, Y. Durachman, S. Harun, A. F. Firmansyah, H. T. Sukmana, and A. Suhaimi, “A rule-based question answering system on relevant documents of Indonesian Quran translation,” in *Proceeding of 2014 International Conference on Cyber and IT Service Management, CITSM 2014*, 2014, pp. 104–107.
- [19] N. Yusliani and A. Purwarianti, “Sistem Question Answering Bahasa Indonesia untuk Pertanyaan Non-Factoid,” *J. Ilmu Komput. dan Inf.*, vol. 4, no. 1, p. 10, 2012.
- [20] M. Murata, S. Tsukawaki, T. Kanamaru, Q. Ma, and H. Isahara, “A system for answering non-factoid Japanese questions by using passage retrieval weighted based on type of answer,” pp. 2–7, 2007.
- [21] A. A. Zulen and A. Purwarianti, “Study and implementation of monolingual approach on Indonesian question answering for factoid and non-factoid question,” in *PACLIC 25 - Proceedings of the 25th Pacific Asia Conference on Language, Information and Computation*, 2011, pp. 622–631.
- [22] A. S. Zharmagambetov and A. A. Pak, “Sentiment analysis of a document using deep learning approach and decision trees,” in *Proceedings of the 2015 12th International Conference on Electronics Computer and Computation, ICECCO 2015*, 2016, pp. 1–4.
- [23] M. Y. Day and Y. Da Lin, “Deep learning for sentiment analysis on google play consumer review,” in *Proceedings of 2017 IEEE International Conference on Information Reuse and Integration, IRI 2017*, 2017, pp. 382–388.
- [24] J. Zhang and C. Zong, “Deep neural networks in machine translation: An overview,” *IEEE Intell. Syst.*, vol. 30, no. 5, pp. 16–25, 2015.
- [25] S. P. Singh, A. Kumar, H. Darbari, L. Singh, A. Rastogi, and S. Jain, “Machine translation using deep learning: An overview,” in *Proceeding of 2017 International Conference on Computer, Communications and Electronics, COMPTHELIX 2017*, 2017, pp. 162–167.
- [26] S. P. Singh, A. Kumar, A. Mangal, and S. Singhal, “Bilingual automatic text summarization using unsupervised deep learning,” in *Proceeding of International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT)*, 2016, pp. 1195–1200.
- [27] C. Yao, J. Shen, and G. Chen, “Automatic document summarization via deep neural networks,” in *Proceedings of 2015 8th International Symposium on Computational Intelligence and Design, ISCID 2015*, 2016, pp. 291–296.
- [28] M. Tan, C. dos Santos, B. Xiang, and B. Zhou, “LSTM-based deep learning models for non-factoid answer selection,” in *Proceeding of the 4th International Conference on Learning Representation (ICLR 2016)*, 2016, no. 1, pp. 1–11.
- [29] L. Yu, K. M. Hermann, P. Blunsom, and S. Pulman, “Deep learning for answer sentence selection,” 2014. [Online]. Available: <https://arxiv.org/abs/1412.1632>.
- [30] T.-Y. Liu, “Learning to rank for information retrieval,” *Found. Trends Inf. Retr.*, vol. 3, no. 3, pp. 225–331, 2009.
- [31] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [32] T. Mikolov, G. Corrado, K. Chen, and J. Dean, “Efficient estimation of word representations in vector space,” in *ICLR*, 2013, pp. 1–12.
- [33] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” in *Proceedings of The 17th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019*, 2019, pp. 4171–4186.
- [34] F. Koto, A. Rahimi, J.H. Lau, and T. Baldwin, “IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP,” in *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020*, 2020, pp. 757–770.

# Assessment of the Healthcare Administration of Senior Citizens from Survey Data using Sentiment Analysis

Ramona Michelle M. Magtangob<sup>1</sup>, Thelma D. Palaoag<sup>2</sup>  
Catanduanes State University Catanduanes, Philippines<sup>1</sup>  
University of the Cordilleras, Baguio City, Philippines<sup>2</sup>

**Abstract**—Healthcare is most frequently used by older people and understanding how they feel about the way healthcare administration gives them the attention and support they need, is crucial to building a healthcare system that is effective in meeting their needs. This study determined the seniors' opinions on the healthcare administration by employing SurveyMonkey, a robust online survey tool as an opinion miner. The study used the Orange application, which made data processing simple, to gauge the seniors' opinions toward healthcare administration by analyzing text sentiment using the VADER Sentiment Analysis, which may distinguish between the polarity of positive, negative, or neutral emotions as well as their intensity. Results showed that the majority of seniors (51.1%) had a negative response to healthcare administration, whereas 47.9% had a neutral response and 1.0% had a positive response. Based on the study, the government should enhance its senior citizens' healthcare services to better satisfy their demands and ensure their happiness. This is clear from the respondents' feedback regarding the services they would like to utilize and how they believe they may be improved. Additionally, the findings provided sufficient information for future consideration to enhance seniors' satisfaction with developmental activities and programs and improve healthcare administration.

**Keywords**—Sentiment analysis; opinion mining; senior citizen; healthcare services

## I. INTRODUCTION

The world's population is aging swiftly, and senior citizens have the fastest rate of population growth. The World Health Organization claims that this trend is anticipated to continue [1]. Health and well-being are crucial for seniors, and the enhancement of their quality of life is a major concern for modern society due to their chronic health difficulties, including the available healthcare services. Due to the growing aging population and increasing life expectancy, seniors need care and access to sophisticated healthcare systems that might enhance their standard of living [2]. Understanding the care and support needs from their perspective and the larger context of their experiences with the services is crucial to successfully addressing those needs [3]. Indeed, we need to reconsider how healthcare administration is being served to senior citizens.

The Philippines is one of the nations with an aging population. As of 2015, 7.4% of the population was over 60 years old, and by 2045, 15.9% of people are expected to be over 60 years old [4]. Given the rapidly aging population, the

complexity of senior citizens' needs, the significance of health information in service delivery, and the difficulties associated with health information in general, it is crucial to ascertain the state of senior citizens' healthcare system to coordinate the country's healthcare services and promote healthy aging [5]. Increasing access to high-quality healthcare is of the utmost importance in emerging countries like the Philippines, and the government needs to be committed to enhancing the healthcare system, especially for vulnerable populations like senior citizens [6].

As the population of seniors increases, particularly in the provinces, their use of health services is becoming more important. Catanduanes is one of the provinces in Region V - Bicol, located in the Southeast part of Luzon in the Philippines. In order to reach its health goals, the provincial government is working hard to achieve health-related goals through its scorecards for planned activities to analyze health issues, however, the prevalence of various diseases continues to rise, despite the provincial government's attempts. This suggests that based on the province's geographic situation and demography, ongoing assessments of health services and inhabitants' health problems should be responsive and appropriate [7], most importantly for senior citizens.

In order to strengthen programs that can help manage the healthcare services of senior citizens, this study collects information on how healthcare services are being administered to them and how they feel about the care they received. By successfully recommending programs that have been proven beneficial and efficient for those in the city and those in the provinces, this study solves information gaps on the need to improve healthcare services given to senior citizens for their health and well-being.

The study will make use of sentiment analysis, a text classification technique that groups texts according to the sentiment of the thoughts they contain. To determine whether a textual analysis is positive, negative, or neutral, it aims to extract the sentiment polarity from the text. Natural language processing (NLP) research in this area is among the most active, with both text and data mining projects being conducted [8][9]. Sentiment analysis has previously been the topic of extensive research and it gives us a method for accurately comprehending sentiments and extrapolating important information from them, it is also a type of text classification that integrates NLP, machine learning, data mining,

information retrieval, and other academic areas [10][11]. A database can be created that contains precise information about healthcare management, including the quality of their services and suggestions for improvements, by efficiently gathering and analyzing the thoughts and opinions of senior citizens.

Using VADER Sentiment Analysis, which is also known as Valence Aware Dictionary for Sentiment Reasoning, this study's overall goal is to evaluate the healthcare administration for senior citizens and develop a methodology that enables them to contribute data that can be utilized to analyze their perceptions, experiences, expectations, and concerns about the healthcare services they received. A profiling and online survey were conducted on the sentiments/opinions of senior citizens towards healthcare administration using SurveyMonkey. Additionally, the Orange Data Mining application tool performed a sentiment analysis evaluation. The sentiment analysis of senior citizens' healthcare administration's online survey was done using the VADER method. Since it was considered to be a time and money-saving approach for minor evaluation efforts, researchers employed the SurveyMonkey software as an assessment tool to gather user comments and opinions [12].

The purpose of this study is to unearth the problem and hopeful insights of the senior citizens, as indicated by the sentiment analysis based on the completed survey. The importance of this study lies in the possibility that it may establish the foundation for the management of senior citizens' healthcare as it discovers opportunities for service enhancement. The researchers' goal in conducting this study is to provide guidance to government policymakers and healthcare administrators on how to prioritize meeting the needs of seniors in order to provide them with high-quality healthcare services. In addition, this may also help the provincial and rural health office to improve its services for older people in terms of health and wellness. By working to make it a reality, we are supporting the UN's third Sustainable Development Goal to promote health and well-being.

## II. RELATED WORKS

### A. Healthcare Services for Senior Citizens

Discussions about how seniors use healthcare services are becoming more and more crucial as the senior population increases. A study being conducted in South Korea aims to create an integrated healthcare service system that is centered on elderly citizens, meeting their needs in daily life and promoting well-being, wellness, and well-dying. A natural structure of regular care, professional care, and rehabilitation for senior members of society in line with the responsibilities of the patients, their families, and caregivers are required for the implementation of the integrated medical care system for elderly users presented in this study [13].

The study [14]'s goal was to determine what senior citizens need from "embedded retirement facilities (ERFs)," multipurpose, and community-based care facilities for the elderly in mainland China. This study is based on questionnaire data collected in northeast China. The findings show that senior citizens' healthcare services are deemed to be the most significant. Senior citizens use community-based facilities, but

decision-makers and facility administrators frequently fail to consider their needs. Seniors in China also tend to be inactive and largely silent in both formal and informal civic involvement because they typically believe that policymakers would take notice of and accommodate their needs.

The purpose of the [15] study is to evaluate older people's well-being to explore whether the data are consistent with previously announced changes in senior treatment in relation to the real resources provided to their patients. The respondents reported being generally satisfied with their lives. The results show that small-town residents felt substantially worse about their quality of life than seniors from large cities. This shows that the healthcare system continues to utterly fail to meet patients' actual demands, particularly in the elderly sector. Being open to a broader discussion about the diverse needs and resources that elderly people in rural and urban areas face is crucial for doing this.

The study of [6] attempts to assess the potential influences on elderly persons' use of healthcare in Davao City, in the Philippines. Various factors were discovered to be significant predictors of healthcare consumption through the use of multiple regression analysis. The findings demonstrated how socioeconomic demographic, personal characteristics and health insurance knowledge affect the way senior citizens use healthcare. By launching health insurance awareness campaigns and creating health-improving initiatives, policymakers and local government organizations may think about enhancing senior citizens' access to healthcare services.

### B. Sentiment Analysis on Healthcare Services

However, in order for various stakeholders to profit from the knowledge gleaned through sentiment analysis, a study by [16] proposes the use of sentiment analysis of hybrid techniques through the development of a module that makes it possible to integrate sentiment analysis functionalities in Web applications related to healthcare at the comment and entity levels. The healthcare industry is one of the least researched industries, however, researchers were aware of a few instances where sentiment analysis was applied to help various industries. Thus, a review of the literature was conducted to examine sentiment analysis utilized in various healthcare settings. The data produced after using sentiment analysis to understand how people feel about various healthcare-related topics enables various stakeholders to take decisions that are advantageous to them. These studies shown allow for validation that the module can be incorporated into Web applications, provide sentiment analysis functionalities to them, and provide various benefits based on the information obtained after using its functionalities without requiring a great deal of effort, like making better decisions or obtaining the reputation and knowing the opinion on aspects related to the healthcare domain.

Due to the vast amount of online information available on healthcare, [17] employed sentiment analysis in this field. They came to the conclusion that sentiment analysis has many advantages, one of which is its ability to use medical data to achieve the greatest results and improve the standard of healthcare. They offer approaches and methods for sentiment analysis employed in the medical field and look for additional

information to help users make the best decision regarding the subject under study.

### C. VADER Sentiment Analysis

VADER is also known as Valence Aware Lexicon and Sentiment Reasoner. The VADER vocabulary was created using conventional sentiment lexicons. Also, this work offers machine learning techniques for sentiment analysis as well as sentiment intensity and orientation lexicons. In order to better understand how the public feels about different entities, this sentiment analysis methods try to identify the feelings of written reviews. Emotions are linked to many characteristics of a product or service as part of the analysis of consumer feedback data. Moreover, VADER sentiment outperformed seven sentiment analysis lexicons, either better or equally well. [18][19][20].

According to a study by [21], VADER maintains and even improves on the advantages of conventional sentiment lexicons like LIWC or Linguistic Inquiry and Word Count: it is larger, yet just as easily examined, understood, and swiftly deployed without requiring substantial learning/training), and it is easily extended. VADER differs from LIWC in that it is both more perceptive of sentiment expressions in social media and more tolerant of generalization to other domains. This can be downloaded and used without charge from the website.

Other research confirms the ease of use of VADER's rule-based sentiment analysis. A compilation of lexical features and their corresponding emotion metrics make up this document. Several guidelines are developed based on the language's grammatical and syntactical usage, and these rules are utilized to assess the text's mood. VADER employs a rule-based method and assigns values to each word in the text in order to consider both the sentiment category and the intensity or strength of the text in addition to the sentiment category. It also performs far faster than machine learning algorithms [22][23].

VADER excels across a range of domain types. Compared to machine learning techniques, VADER has a number of advantages. It is firstly quick and computationally effective. The second advantage is that the terminology and regulations of the VADER are clear and not hidden. Because of this, VADER is easy to understand, build upon, and alter. By setting the threshold at 0.05, VADER is a preferable option if processing the sentiment quickly and if it was the only thing that had been planned. VADER also adheres to grammatical and syntactical rules for expressing and highlighting sentiment intensity. VADER outperforms Text blob and NLTK sentiment analysis technologies in terms of performance. [24].

According to empirical findings, the technique utilized is the best technique for ranking many choices. Additionally, users of the healthcare sector's decision-making processes and healthcare providers' goals for quality improvement can both benefit from ranking information.

## III. METHODOLOGY

This study is divided into four main phases, the first of which is the actual data gathering from senior citizens whose opinions will be utilized to determine the study's analysis of the government healthcare system. As can be seen, the data will be cleaned using a preprocessing procedure in the phase after which it will be tagged with the necessary sentiment labels using the VADER method. The orange data mining application will be utilized in this step. Preprocessing entails tidying up the text, making it all lowercase to make it easier to read, and removing any affixes from words to get them down to their root forms.

To understand more about the primary healthcare administration for seniors, particularly in the province, the researcher conducted a survey on senior citizens in Virac, Catanduanes. The study was carried out in the barangay San Isidro Village since it has the highest number of senior residents in the municipality. A total of 694 responses were gathered using the SurveyMonkey software which is an online survey tool. This online survey software makes designing and managing reliable online surveys easy. It is also a highly effective and well-known online platform.

This research demonstrates how to retrieve the required data from its source using the Orange Data Mining application and how to use tools to carry out text mining operations. The procedures, particularly when applying the classification technique using the VADER method of Sentiment Analysis of Orange Application, are shown in Fig. 1.

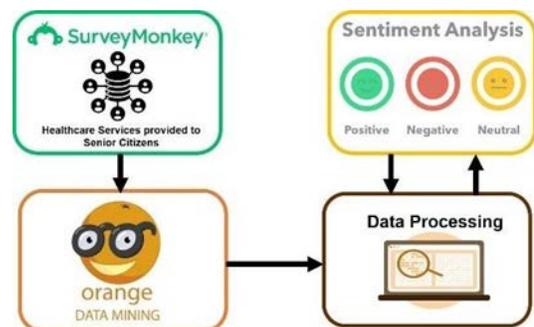


Fig. 1. Using orange data mining application for the sentiment analysis

### A. Data Collection and Gathering

To understand more about the primary healthcare administration for seniors, particularly in the province, the researcher conducted a survey on senior citizens in Virac, Catanduanes. The study was carried out in the barangay San Isidro Village since it has the highest number of senior residents in the municipality. A total of 694 responses were gathered using the SurveyMonkey software which is an online survey tool. This online survey software makes designing and managing reliable online surveys easy. It is also a highly effective and well-known online platform.

The survey provided was composed of three questions, the first was the name of the respondent, which was also optional, their age, and the third question asked about senior citizens' opinions of the healthcare services provided.

Fig. 2 shows the 694 data collected from the senior citizens' responses using the SurveyMonkey software. The responses were gathered and downloaded in a comma-separated (CSV) file that would be utilized in the following step.

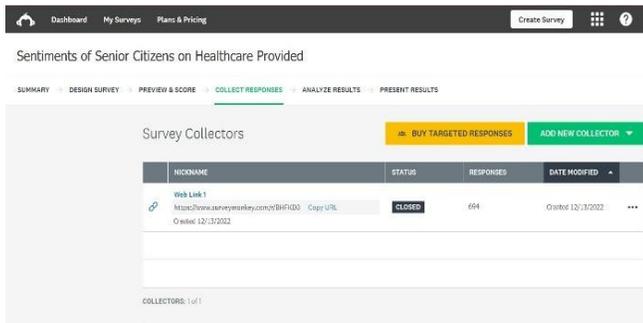


Fig. 2. Collected data using SurveyMonkey tool

**B. Data Processing**

Using the Orange Data Mining Application, the survey text corpus from the SurveyMonkey responses was imported and processed. The corpus goes through pre-processing to ensure that it was converted to lowercase and removed unwanted words. To verify the outcomes, a word cloud was additionally attached to the pre-process text widget.

Fig. 3 depicts the loading of SurveyMonkey survey text into the corpus, and Fig. 4 depicts the pre-processing of corpus texts acquired on senior citizens' opinions of the healthcare services offered to them by the governments.

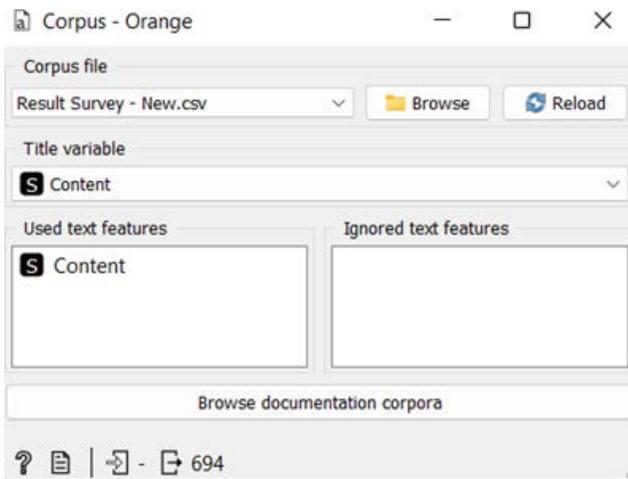


Fig. 3. SurveyMonkey text responses gathered

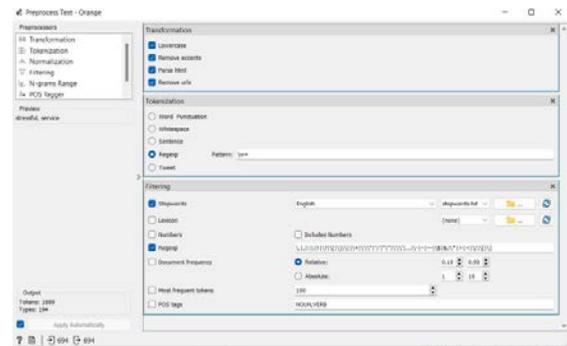


Fig. 4. Corpus texts' pre-processing

**C. Sentiment Extraction**

The analysis of the survey data, also known as sentiment extraction, is another important stage. This analysis used topic modeling and VADER sentiment analysis. In this instance, the comma-separated (CSV) file, which was processed prior to sentiment extraction, is extracted to obtain the widget. The CSV file was turned into the corpus by selecting text characteristics from the dialog box, which was then processed.

The sentiments taken from the SurveyMonkey tool, which had previously applied VADER sentiment analysis, are shown in Fig. 5. Additionally, it displays the text's measured emotions, whether they are neutral, positive, or negative.

id	Content	Name	Age	pos	neg	neu	compound
1	stressful service	lino	63	0	0.767	0.233	-0.5106
2	cause anxiety	maria	60	0	0.63	0.37	-0.1779
3	cause anxiety	julia	67	0	0.63	0.37	-0.1779
4	bad experience...	tere...	71	0	0.467	0.533	-0.5423
5	dissappointed ...	am...	64	0	0	1	0
6	not good	jose	65	0	0.706	0.294	-0.3412
7	cause anxiety	julian	62	0	0.63	0.37	-0.1779
8	bad service	juana	70	0	0.778	0.222	-0.5423
9	bias treatment	elena	69	0	0.583	0.417	-0.1027
10	stressful service	asu...	63	0	0.767	0.233	-0.5106
11	poor service	mar...	67	0	0.756	0.244	-0.4767
12	noot good	teresa	66	0.744	0	0.256	0.4404
13	burdensome ser...	julia...	68	0	0.737	0.263	-0.4215
14	stressful services	anna	61	0	0.767	0.233	-0.5106
15	delayed actions...	luisa	60	0	0.241	0.759	-0.2263
16	depriving qualif...	losito	73	0	0.6	0.4	-0.4588

Fig. 5. Sentiments extracted from SurveyMonkey tool

**D. Using Sentiment Analysis**

Sentiment analysis was used in this study to determine whether the content of survey responses contained neutral, positive, or negative emotions. This kind of text analytics employs both machine learning and natural language processing (NLP). Fig. 6 shows the steps taken to achieve sentiment analysis for the survey of senior citizens. From the survey text corpus from responses from SurveyMonkey to the pre-processing of the text and word cloud, followed by the sentiment analysis using the VADER method, to repeating the pre-processing texts to ensure that there are no unwanted words included, and choosing the necessary columns to include in the result, to providing the corpus viewer to view the results of the sentiment analysis using the VADER method.



Ineffective health monitoring and counseling, technology to assist with procedures, inadequate service, and mobility issues were some factors that sparked criticism. These factors all had an impact on how they perceived the services offered to them.

Significant relationships were also discovered between respondents' perceptions of the standard of healthcare. Based on the study, the government should enhance its senior citizens' healthcare services to better satisfy their demands and ensure their happiness. This is clear from the respondents' feedback regarding the services they would like to utilize and how they believe they may be improved. However, this study has important limitations that could encourage additional development and research. Seniors' views expressed on social media and the Internet in the provinces may one day be evaluated for several reasons. A better healthcare system may be created in response to unfavorable comments in order to effectively address and support seniors' concerns and improve the government's provision of healthcare services. This will assist the government in creating a plan to satisfy the demands of senior citizens while providing a suitable healthcare system. This will also show how important it is to have a reliable healthcare system across the country, not only in urban and rural areas. This will guarantee that senior folks are treated with the respect they deserve and that information concerning their health is immediately updated.

#### ACKNOWLEDGMENT

The authors would like to express their gratitude to everyone who took part in this study, gave their time and expertise to provide the essential data, and took the time to carefully respond to the survey questions from the beginning of the project's conceptualization through its completion.

#### REFERENCES

- [1] "Ageing," 12 May 2022. [Online] Available: <https://www.who.int/news-room/facts-in-pictures/detail/ageing>
- [2] Etemad-Sajadi, R., & Gomes Dos Santos, G.(2019). Senior citizens' acceptance of connected health technologies in their homes. *International Journal of Health Care Quality Assurance*, 32(8), 1162– 1174. doi:10.1108/ijhcqa-10-2018-0240
- [3] Abdi, S., Spann, A., Borilovic, J., Witte, L., & Hawley, M. (2019) Understanding the care and support needs of older people: a scoping review and categorization using the WHO international classification of functioning, disability, and health framework (ICF). *BMC Geriatr* 19, 195. <https://doi.org/10.1186/s12877-019-1189-9>
- [4] Cruz, G. T., Saito, Y., Cruz, C. J. P., & Paguirigan, M. R. B. (2019). The 2018 Longitudinal Study of Ageing and Health in the Philippines. *Ageing and Health in the Philippines*; Cruz, GT, Cruz, JCP, Saito, Y., Eds.
- [5] Garcia AP, De La Vega SF, Mercado SP. (2022) Health Information Systems for Older Persons in Select Government Tertiary Hospitals and Health Centers in the Philippines: Cross-sectional Study. *J Med Internet Res*. 2022 Feb 14;24(2):e29541. doi: 10.2196/29541. PMID: 35156927; PMCID: PMC8887638.
- [6] Alipio, M., & Pregoner, J. D. (2020). Determinants of Healthcare Utilization among Senior Citizens in Davao City, Philippines. *JPAIR Multidisciplinary Research*, 39(1), 50–65. <https://doi.org/10.7719/jpair.v39i1.759>
- [7] Panti, M. B. (2020). Health Status and Programs Implemented by the Provincial Government of Catanduanes in Bicol Region, Philippines. *Open Journal of Social Sciences*, 8(5), 419-431.
- [8] García-Díaz, J. A., Cánovas-García, M., & Valencia-García, R. (2020). Ontology-driven aspect-based sentiment analysis classification: An infodemiological case study regarding infectious diseases in Latin America. *Future Generation Computer Systems*. doi:10.1016/j.future.2020.06.019
- [9] Devika, M. D., Sunitha, C., & Ganesh, A. (2016). Sentiment Analysis: A Comparative Study on Different Approaches. *Procedia Computer Science*, 87, 44– 49. doi:10.1016/j.procs.2016.05.124
- [10] Xu, G., Meng, Y., Qiu, X., Yu, Z., & Wu, X. (2019). Sentiment Analysis of Comment Texts Based on BiLSTM. *IEEE Access*, 7, 51522–51532. doi:10.1109/access.2019.2909919
- [11] Abirami, A. M., & Askarunisa, A. (2017). Sentiment analysis model to emphasize the impact of online reviews in healthcare industry. *Online Information Review*, 41(4), 471– 486. doi:10.1108/oir-08-2015-0289
- [12] Al-Shabi, M.A. (2020) Evaluating the performance of the most important Lexicons used to Sentiment analysis and opinions Mining. *Int. J. Comput. Sci. Netw. Secur.*, 20, 51–57.
- [13] Sourav De, Sandip Dey, Surbhi Bhatia, Siddhartha Bhattacharyya (2022) Chapter 1 - An introduction to data mining in social networks, In *Hybrid Computational Intelligence for Pattern Analysis, Advanced Data Mining Tools and Methods for Social Computing*, Academic Press, Pages 1-25, ISBN 9780323857086, <https://doi.org/10.1016/B978-0-32-385708-6.00008-4>.
- [14] Xiang, L., Yu, A. T. W., Tan, Y., Shan, X., & Shen, Q. (2019). Senior citizens' requirements of services provided by community-based care facilities: a China study. *Facilities*, 38(1/2), 52–71. doi:10.1108/f-02-2019-0023
- [15] Błażnio A, Buliński L. (2019) Wellbeing and older adults in primary health care in Poland. *Ann Agric Environ Med*; 26(1): 55–61. doi: 10.26444/aaem/85711
- [16] Ramírez-Tinoco, F. J., Alor-Hernández, G., Sánchez-Cervantes, J. L., Salas-Zárate, M. del P., & Valencia-García, R. (2019). Use of Sentiment Analysis Techniques in Healthcare Domain. *Current Trends in Semantic Web Technologies: Theory and Practice*, 189–212. doi:10.1007/978-3-030-06149-4\_8.
- [17] Abualigah, Laith & Alfar, Hamza & Shehab, Mohammad & Abu Hussein, Alhareth. (2020). Sentiment Analysis in Healthcare: A Brief Review. 10.1007/978-3-030-34614-0\_7.
- [18] Ramyasri, V & Niharika, Ch & Maneesh, K & Ismail, Mohammed. (2019). Sentiment Analysis of Patients' Opinions in Healthcare using Lexicon-based Method. *International Journal of Engineering and Advanced Technology*. 9. 2249-8958. 10.35940/ijeat.A2141.109119.
- [19] Bansal, A., Kumar, N. (2022) Aspect-Based Sentiment Analysis Using Attribute Extraction of Hospital Reviews. *New Gener. Comput.* 40, 941–960. <https://doi.org/10.1007/s00354-021-00141-3>
- [20] Borg, A., & Boldt, M. (2020). Using VADER Sentiment and SVM for Predicting Customer Response Sentiment. *Expert Systems with Applications*, 113746. doi:10.1016/j.eswa.2020.113746
- [21] Kim, YS., Lee, J., Moon, Y. et al. (2020) Development of a senior-specific, citizen-oriented healthcare service system in South Korea based on the Canadian 48/6 model of care. *BMC Geriatr* 20, 32. <https://doi.org/10.1186/s12877-019-1397-3>
- [22] Pano, Toni, and Rasha Kashef (2020). "A Complete VADER-Based Sentiment Analysis of Bitcoin (BTC) Tweets during the Era of COVID-19" *Big Data and Cognitive Computing* 4, no. 4: 33. <https://doi.org/10.3390/bdcc4040033>
- [23] Siddhaling Urologin, (2018) "Sentiment Analysis, Visualization, and Classification of Summarized News Articles: A Novel Approach" *International Journal of Advanced Computer Science and Applications(IJACSA)*, 9(8), <http://dx.doi.org/10.14569/IJACSA.2018.090878>
- [24] Bonta, V., & Janardhan, N. K. N. (2019). A comprehensive study on lexicon based approaches for sentiment analysis. *Asian Journal of Computer Science and Technology*, 8(S2), 1-6.

# Hierarchical Pretrained Deep Learning Features for the Breast Cancer Classification

Abeer S. Alsheddi

Computer Science Department, Imam Mohammad Ibn Saud Islamic University, Riyadh, Saudi Arabia

**Abstract**—Breast cancer is a common and fatal disease among women worldwide. Accurately and early diagnosing of breast cancer plays a pivotal role in improving the prognosis of patients. Recently, advanced techniques of artificial intelligence and natural image classification have been used for the breast cancer image classification task and have become a hot topic for research in machine learning. This paper proposes a fully automatic computerized method for breast cancer classification using two well-established pretrained CNN models, namely VGG16 and ResNet50. Next, the feature extraction process is used to extract features in a hierarchical manner to train a support vector machine classifier. Evaluating the proposed model shows achieving 92% accuracy. In addition, this paper investigates the effect of different factors, highlights its findings, and provides future directions for the research to develop more advanced models.

**Keywords**—Feature extraction; CNN models; Pretrained models; breast cancer classification

## I. INTRODUCTION

Breast cancer is the most common form of cancer in women, and invasive ductal carcinoma (IDC) is the most common form of breast cancer [1]. It is frequently occurring and increasingly fatal. Usually, a biopsy is taken from patients and then a pathologist must decide whether they have breast cancer or not. Manual diagnosis from slides is time-consuming and the decision itself depends on the expertise of the pathologist and their equipment. Image Processing and Deep Learning can be used to create models that complement doctors by automating and speeding up diagnosis to save time and minimize errors in detecting breast cancer. This problem is not new these days, one of the earliest DNN applications was on breast cancer images [2].

1) *Research problem*: in this paper, the main goal is to classify breast cancer images in the form of images into binary classification: IDC and non-IDC. To formulate the problem, let  $x$  be a 2D image that belongs to  $R^{m \times n}$  where  $R$  is the space of 2D images with width  $m$  and height  $n$ , and let  $Y = \{0, 1\}$  where 0 indicates no breast cancer (non-IDC) and 1 indicates to indicating breast cancer was found in the image (IDC). Then the problem of the breast cancer classification is to model a mapping  $f$  from  $R^{m \times n}$  to  $Y$ , such that,

$$f(\cdot, \theta): R^{m \times n} \rightarrow Y \quad (1)$$

where any value of  $x$  will be mapped to  $y$ ,  $f(x, \theta) = y$ , for any  $x \in R^{m \times n}$  and  $y \in Y$ .

2) *Research objective*: the main contribution of the proposed model in this paper is combining between two tasks; First: using multiple pretrained models to extract features; second: using a feature hierarchy concept during extracting the features. To achieve this contribution, two pretrained VGG16 [3] and ResNet50 [4] models that have excellent classification performance for natural image classification in the Image Large Scale Visual Recognition Challenges, are used to extract activations of different five convolution layers from each model, ten layers in total. Next, the features are reduced using pooling operations to be  $6 \times 6 \times 5$  at each pretrained model and then concatenated them to result a  $6 \times 6 \times 10$  layer. Finally, the resulting features feed to the support-vector machines (SVM) classifier [5]. The results show that the combined feature hierarchy from two pretrained models gets 92% accuracy higher than using a pretrained model individually.

The rest of the paper is recognized as the following. Section II reviews some of the state-of-art in the problem. Next, Section III provides the proposed method and Section IV presents its results and discusses them. Finally, Section V concludes the main results and provides some future directions.

## II. RELATED WORK

This section reviews some of the research that works on the breast cancer classification problem and gives information about the used methods. Table I summarizes the methods of these related works. Most research recently uses different techniques to deal with features and then applies the classification on these features instead of classifying the whole image using ANN.

TABLE I. METHODS SUMMARIZATION IN THE RELATED WORK

Types	Methods	Research
Feature extraction from pretrained models	Fuses activations from FC of three pretrained models	[6]
	Fuses the results of the classifiers and extracted activations from each FC in pretrained model	[7]
Transfer parameters	Fine-tuned pre-trained models with using logistic regression classifier	[8]
	Fine-tuned modified AlexNet model	[9]
Unsupervised learning with segmentation step	Selecting features using GA algorithm and CNN model for classification.	[10]
	Clustering using Lloyd's algorithm [11] for and CNN model for classification.	[12]
Feature selection	Feature selection using mRMR algorithm with 4 classifier: SVM, Naïve Bays, Function tree and End Meta	[13]

The researchers usually try to find a proper feature representation of images to train their model. One major way is using a pretrained DCNN to extract an image activation as its features. Several pre-trained models were used for this process. The proposed model in [6] extracts features from fully connected layers of three models, namely GoogleNet [14], VGG [3], ResNet [4] models. Then classifier was trained on these features. The accuracies of these three models individually were 93.5%, 94.15%, and 94.35% respectively, while the combination between them achieves 97.525% accuracy. Thus, fusing the features from different models leads to better classification compared to extracting from a single model. The other research [7] uses three pre-trained models, namely AlexNet [15], VGG16 [3] and ResNet-18 [4] to extract features and then uses them to train three SVM classifiers, one classifier for each pretrained model. Instead of combining the features, the research fuses the three results from each classifier by calculating the average and combining the probabilities for fusion to obtain the final decision score. It measures the performance using the receiver operating characteristics curve (AUC) and it achieves 83.83% and 97.55% for two different datasets. However, for both research, [6] and [7], FC layers contain usually a high number of activations than convolution layers which will consume more computation cost.

In transfer parameters, the learning assumes that the two models share some parameters that can learn effectively. The research [8] analyzes different pretrained models VGG [3] and ResNet [4] considering all activations values of the convolution layers without considering the fully-connected layers and using the same strategy of the previous research [16] but without using one pre-trained model, namely AlexNet [15] as done [16]. Next, the logistic regression classifier is to decide the predicted class. As a result, a fine-tuned pre-trained VGG16 achieves the best performance at 92.60% accuracy. The other research [9] uses the same strategy as AlexNet [15]. The authors adapted the AlexNet with some modifications in its architecture related to the normalization process and type of the activation function. These modifications provide different proposed models. Then they applied fine-tuned processing on the models and achieved individual model ranges between 75% and 77%, while the combining model 84% accuracy.

The other research [10] uses unsupervised learning to implement its model. The proposed model is based on k-mean algorithm [17] and the probabilistic model (GMM). The proposed model first finds the region of interest (ROI) and then applied the feature selection using genetic algorithms (GA) [18]. Next, the model applies the CNN algorithm to find out better results. The resulted accuracy achieved 95.8%. The other research [12] also uses the segmentation step before classification using Lloyd's algorithm [11] for clustering and CNN for classification. A 96% accuracy was achieved by the proposed methods. However, in both the previous research, the authors did not mention exactly the proposed CNN model that was used.

The feature selection is also used as a preprocess of classification in a hybrid approach [13]. It uses a minimum redundancy feature selection (mRMR) algorithm [19] to effectively identify object properties and narrow down their relevance and then can predict breast cancer. The proposed

approach uses four classifiers SVM, Naïve Bays, Function tree and End Meta to find out the best performance. The result shows that SVM outperforms at 99% accuracy on average by combining it with MRMR algorithm. However, the feature selection process may not be enough for training with large datasets without dealing with deep learning models. This proposed approach is not the only research that claims the outperformance of the SVM classifier. A number of research early and recently [20] [21] [22] [23] [24] compare different classifiers and reach the same result, such as the study in [25] focuses to compare random forest and SVM classifiers for breast cancer classification and claimed that the highest accuracies 95% is for SVM.

On the other hand, some research not related to breast cancer classification uses a feature hierarchy to represent the images in CNNs. The research [26] studies real-world video sequences. It uses different hierarchical features of convolutional layers in CNNs to deal with features at early layers that keep more fine-grained spatial details and are useful for localization. It claimed that dealing with multiple layers of CNN features to get better performance for learning video features and visual tracking.

In the end, the result shows that the concept of transfer learning can be successfully applied to the breast classification domain. The activations of the source model can be used as features in the target model in the breast classification domain with less implementation cost, i.e., using pretrained models instead of training from scratch. Moreover, the combination of the features from different neural networks improves the accuracy of the classifiers.

In addition, a common way in the previous research using the activations of the fully connected (FC) layer as features. However, FC layers contain a higher number of activations than convolution layers which consume more computation cost. Also, extracting features from different layers leads to a better performance in learning video features. Moreover, most of the previous research used accuracy as a metric to evaluate their proposed models.

### III. PROPOSED METHOD

The idea of the proposed solution is image classification by extracting a feature hierarchy from pretrained CNN models and then feeding it into a classifier instead of using the whole images as inputs to that classifier. Extracting features from different layers of a single network is shown to lead to better performance in previous research working on learning video features [26].

To build the proposed model, two sub-models are constructed, one for extracting features and the other for classification. Writing  $f_m$  for the final proposed model that its form maintained under composition  $f_c$  and  $f_t$ ,  $f_c$  for the classifier, and  $f_i$  for the features by

$$f_m = f_c(f_i(x), y) \quad (2)$$

where  $x$  is called the input images and  $y$  is the true class of the input images with two possible values 0 or 1, where 0 indicates to no breast cancer (non-IDC) and 1 indicates to breast cancer was found in the image  $x$  (IDC). Exactly, each

input  $x_i$ , where  $x_i \in \mathbb{R}^{m \times n}$  will be mapped to  $y_i$ , where  $y_i \in \{0,1\}$ . Fig. 1 displays the general diagram of the proposed method. While each component of the composition will be described in the following subsections supported by described figures.

A. FT: Feature Extraction Phase

Feeding the whole images into a classifier needs to extract features manually an extremely time-consuming process and needs a strong knowledge of the domain. Also, converting 2D images to 1D vectors increases the number of trainable parameters exponentially and it significantly can increase the chance of overfitting especially if the size of a dataset is less than the number of learnable parameters. Thus, a CNN model is used in this proposed model for extracting the features.

Pretrained CNN model is decided to be used because the process of training networks with a large number of parameters is time- and resource-consuming. Thus, two pretrained models VGG16 and ResNet50 are used in this paper which they are used previously on a similar domain [6] [7].

1) VGG16 model is a type of CNN Architecture proposed by Visual Geometry Group (VGG), Oxford University [3]. Using VGG16 with 16 learnable layers regarding the depth which is larger than 8 layers in AlexNet [15], as an example, gives important for achieving high performance [15]. Moreover, VGG16 shows excellent classification performance for different previous works natural image classification in the Image Large Scale Visual Recognition Challenges [27] and for different previous works [6] [7] [28].

VGG model contains 16 learnable layers separated into five groups where each group ends with a pooling layer. In this proposed model, a pretrained VGG model is used with the input size differs from the default size in VGG16. The input size is equal to 50x50 pixels to be the same size as the input dataset and the three fully connected layers are removed. Fig. 2 (a) is zoomed a portion of Fig. 1 that presents the feature extraction process in VGG16 model. The resulting sizes are presented on Table II after applying the following steps:

a) Creating a VGG16 model without the three fully connected layers due to the purpose of using VGG16 model as a features generator from its intermediate layers, not as a classifier.

b) Feeding the input into the VGG16 model and extracting feature maps at five intermediate layers. The layers are

c) The last layer (pooling layer) in each group, the red layer in Fig. 2 (a). Hence, five layers generate five different blocks of feature maps with different shapes.

d) To combine extracted feature maps from the previous step in a specific axis, the layers must have the same dimension on them. To unify the size to be the same 6x6 as the size of width and high, up/down sampling operations are applied. The down-sampling operation is applied to the first and the second extracted layers using the max pooling layer. The third layer is already having the same required size, so it does not need to change. The up-sampling operation is applied to the fourth and the fifth extracted layers using the transpose convolutional

layer that performs an inverse convolution operation. More detail about the values of their hyper-parameters is detailed in Section 4.1.

e) Each layer has a high number of channels which will increase the computation time. At the same time, the activated region of a channel is semantically meaningful and serves a similar role as the feature detectors to identify different features present in an image [29]. Thus, max pooling over the depth operation is applied to extract the maximum value of activation in a specific location (receptive field) among all channels and decrease the number of channels to only one channel. It is noteworthy that the utilized up-sampling method returns one channel by default. Thus, we can remove this step from the up-sampling layers.

f) Concatenating the five resulted from layers of the previous step on the depth axis to get one 6x6x5 layer. These feature maps will be concatenated with the resulting layer from the ResNet50 model which will be described in the next subsection.

2) ResNet50 model [4] consists of 48 convolution layers along with one max pooling and one average pool layer. The model has two types of connections: Identity connections between every two convolution layers and skip connections between some of them. The skip connections help to solve the vanishing gradient problem by allowing for the gradient to flow through these shortcut paths. Thus, it enables CNN models to get deeper and deeper without decreasing the accuracy by adding more layers to the network.

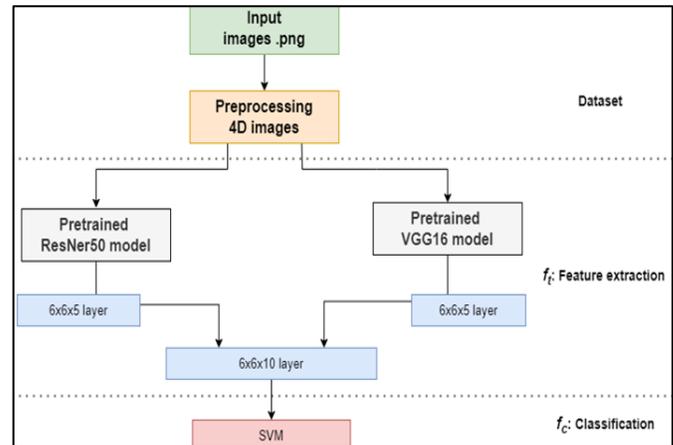


Fig. 1. Diagram of the proposed method

TABLE II. THE SIZES OF RESULTED FEATURES FROM VGG16 MODEL

Layer numbering	Extracted from Model	After up/down sampling	After depth pooling
1	25x25x64	6x6x64	6x6x1
2	12x12x128	6x6x128	6x6x1
3	6x6x256	-	6x6x1
4	3x3x512	6x6x1	6x6x1
5	1x1x512	6x6x1	6x6x1

In this proposed model, a pretrained ResNet model is constructed with the same VGG settings. Fig. 2 (b) presents the feature extraction process and the resulting dimension is presented on Table III applying the following steps:

a) Creating a ResNet model without the fully connected layer due to the purpose of using ResNet model.

b) Feeding the input into the ResNet model and extracting feature maps at the last layer in each group. In total, there are five layers that generate five different blocks of feature maps.

c) The width and high are unified to be 6x6 using up/down sampling operations. The down-sampling operation is applied to the first three layers using the max pooling layer. The up-sampling operation is applied to the fourth and the fifth extracted layers using the transpose convolutional layer.

d) Decrease the number of channels to one channel using the max pooling over the depth operation.

e) Concatenating the five resulted from layers on depth axis to be one 6x6x5 layer.

3) *VGG and ResNet combination.* After building the two models separately, the resulting layer of each model is 6x6x5 layer. The last step in the extracting phase is to concatenate these two layers on the depth axis to be 6x6x10 layers as Fig. 1 shows.

TABLE III. THE SIZES OF RESULTED FEATURES FROM RESNET50 MODEL

Layer numbering	Extracted from Model	After up/down sampling	After depth pooling
1	13x13x64	6x6x64	6x6x1
2	13x13x256	6x6x256	6x6x1
3	7x7x512	6x6x512	6x6x1
4	4x4x1024	6x6x1	6x6x1
5	2x2x2048	6x6x1	6x6x1

### B. Classification Phase

The main goal of this paper is to classify breast cancer in the form of 2D images into binary classification: IDC and not IDC. The resulting features from the pretrained models along with the corresponding labels (i.e., IDC or non-IDC) are then used to train binary non-linear SVM classifier. The SVM classifier performs a good result in different works of breast cancer classification [19] [20] [21] [22] [23] [24] [25].

In SVM implementation, feature scaling is a crucial step because the methodology of SVM considers the distances among inputs to select the maximum decision boundary. This distance is surely different for non-scaled and scaled cases. Thus, the scaled step is applied using standardized features [30] with a mean equal to zero and standard deviation equal to one,

$$z = \frac{x - \mu}{\sigma} \quad (3)$$

where  $x$  is the concatenated feature,  $\mu$  is the mean and  $\sigma$  is the standard deviation of these features. This step makes the features fall in a small range and leads to faster convergence in fewer iterations and then better performance [31].

## IV. RESULTS AND DISCUSSION

This section provides in detail the implementation of the proposed model and presents the results along with discussing it.

### A. Implementation

1) *Pretrained models construction.* The pretrained models are constructed using TensorFlow-Keras package with the same weights pretraining on ImageNet dataset [32] without changing or learning any weight. Thus, all convolution layers of the pretrained models are frozen. Moreover, the input size is equal to 50x50 pixels to be the same size as the input dataset. Although the default input size in the two pretrained models VGG16 and ResNet50 is 244x244 pixels, this proposed method discards the classification part with the fully connected layers to allow any input size. Table IV shows the value for each hyper-parameter in the construction.

2) *Extracting features phase.* To extract a feature hierarchy, ten temporary small models are constructed, five models for each pretrained model. Each small model is prepared to take the inputs equal to the input of the pretrained model and produce a block of feature maps as the output, which are used as features. Recall that the output layers of the small models are different regarding to producing five different layers in each pretrained model. Moreover, to unify the shape of the feature maps, different pooling layers in the Keras package are applied. In case of the layer size greater than 6x6, the down-sampling operation using the method *MaxPooling2D()* is applied, or in case of the layer size is less than 6x6, apply the transpose convolutional layer using the method *Conv2DTranspose()*. Moreover, the depth pooling operation is applied by the method *reduce-max()* to get a maximum element across a specific axis, here the depth. The assigned values for each parameter in the methods are presented on Table V. After unifying the shape of all feature maps, the method *concatenate()* is applied to concatenate all blocks of the feature maps among the depth axis.

3) *Classifier.* To can feed the features into the classifier, the array of the features must be reshaped to be 2D array with the number of inputs as the row and multiple of the 6x6x10 as the columns using the method *reshape()* in tensorflow package. The next step is applying the standardized features using the methods *StandardScaler()* and *transform()* in sklearn package.

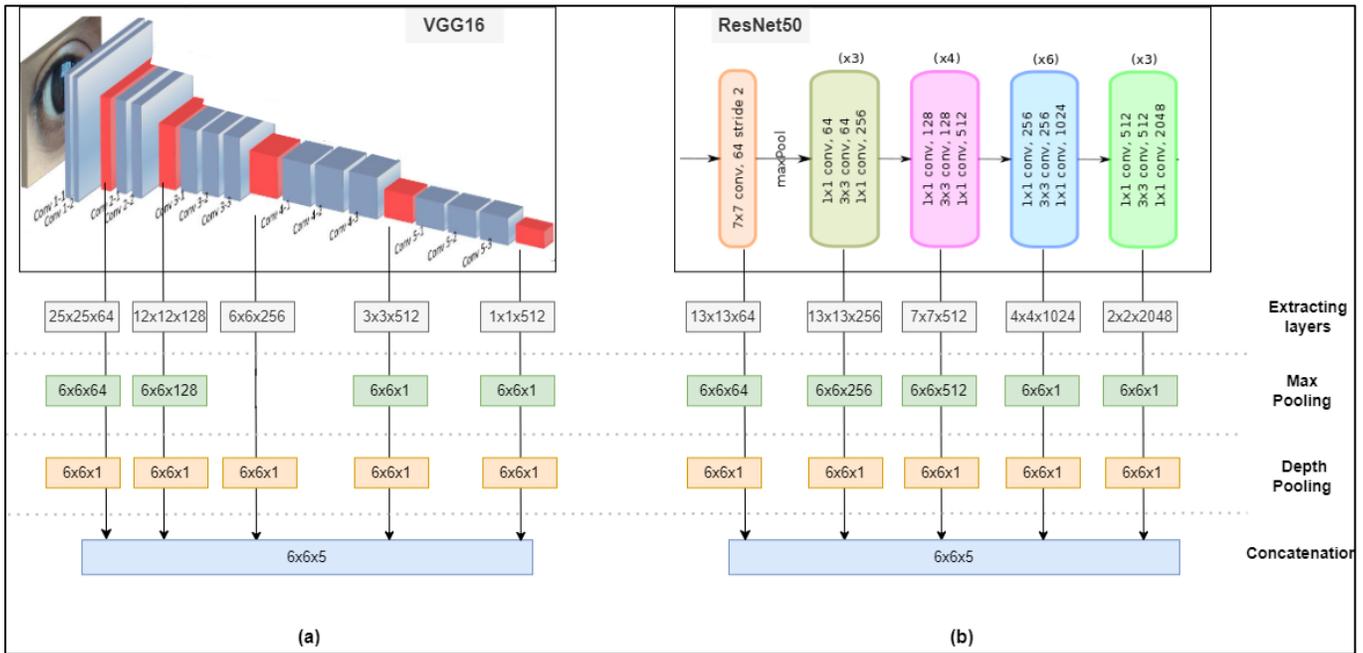


Fig. 2. Diagram of the pretrained (a) VGG16 (b) ResNet50 models.

After that, the non-linear SVM classifier is constructed from sklearn package using different methods. The method *SVC()* constructs the classifier by adjusting three hyper-parameters: regularization parameter *C* for giving a different level of regularization, the kernel parameter for enabling SVM to solve nonlinear classification problems when the inputs cannot be separated linearly, and the gamma parameter for considering as spreading of the inputs that are selected by SVM as support vectors and therefore affect the decision region. When the value of gamma is low, the curve of the decision boundary is very low and thus the decision region is very broad and vice versa. Different values are assigned to these three parameters to estimate the best values. Table VI presents the suggested values for each parameter. The method *GridSearchCV()* helps to loop through the three parameters and fit SVM classifier on the training set to select the best values. The best value is optimized by the cross-validation splitting parameter *CV*.

To complete training SVM classifier, the number of maximum iterations is fixed to 30,000 iterations because the convergence warning appears due to convergence issues. The other solution to overcome this issue is using standardizing features that helps to reach the convergence state faster.

TABLE IV. THE HYPER-PARAMETERS OF CONSTRUCTING PRETRAINED MODELS

Parameters	Values
weights	imagenet
Layer trainable	Fasle (freeze)
Include top (FC)	False
Input size	50 x 50 pixcel
Parameters	Values

TABLE V. THE HYPER-PARAMETERS OF EXTRACTING FEATURE PHASE

Up/down sampling	Layers numbering	VGG16	ResNet50
Down	1	Pooling size= (4, 4)	Pooling size= (2, 2)
	2	Pooling size= (2, 2)	Pooling size= (2, 2)
	3	-	Pooling size= (2, 2) Strides = (1,1) with padding
Up	4	kernel_size = (2,2) Strides = (2,2)	kernel_size = (3,3) Strides = (1,1) with padding
	5	kernel_size = (6,6) Strides = (2,2)	kernel_size = (5,5) Strides = (1,1) with padding

TABLE VI. THE HYPER-PARAMETERS OF EXTRACTING SVM

Parameters	Values
Max iteration	30,000
C	[0.1,1, 5, 10]
kernel	Radial basis function (rbf), Polynomial kernel, sigmoid
gamma	[1,0.1,0.01,0.001]
Parameters	values

### B. Evaluation

1) *Cross-validation*. The cross-validation evaluates a classifier's performance by dividing the dataset into *k* parts. *K* is equal to 10 in this paper which is called 10-fold cross-validation. Thus, each image in this dataset will be used 9 times for training and once for testing. This validation then calculates the average between them to evaluate the classifier's performance. Thus, to evaluate the performance of the trained

classifiers, the cross-validation splitting parameter CV of the evaluation methods is assigned to 10 as 10-fold cross-validation.

2) *Performance metrics.* The Metric is accuracy as it is used in most of the previous works in Section II. Accuracy measures how many IDC and non-IDC images are classified correctly among all classifications. It shows overall how is the classifier classified correctly. Calculating the accuracy of the training set as an average over 10-cross-validation folds. Especially, the experiments are made in three cases for three SVM classifiers. Each SVM classifier is related to one of the following models: using only VGG16 model, only ResNet50, and using the combination of both models.

3) *Test platform.* The experiments are concurred using a personal laptop. However, the GPU in the laptop is not supported by python. Some of the tasks then run in long execution times and the memory of the laptop may not be enough. Thus, I have moved to use Google Colab Pro due to some commands could not be run using a free version of Google Colab.

4) *About the dataset.* The used dataset of the breast-cancer-image-classification is available in [33]. Fig. 3 shows the distribution of the dataset. The original images are for 279 patients with a small number of images scanned at 40x. However, overfitting is highly likely. Then, 50x50 patches were extracted including 198,738 negative examples (i.e., no breast cancer) and 78,786 positive examples (i.e., indicating breast cancer was found in the patch). Thus, the available dataset contains 277,524 patches in total. According to the figure, there is clearly an imbalance in the class data with over two times the number of negative data points than positive data points. However, in this work, the loading step, which loads the whole dataset into a programming notebook, has caused a crash multiple times after running the code in hours because of the available RAM space in Google Colab Pro. This leads to using a part of the dataset in the experiments with keeping the same percentage of imbalance in the class data.

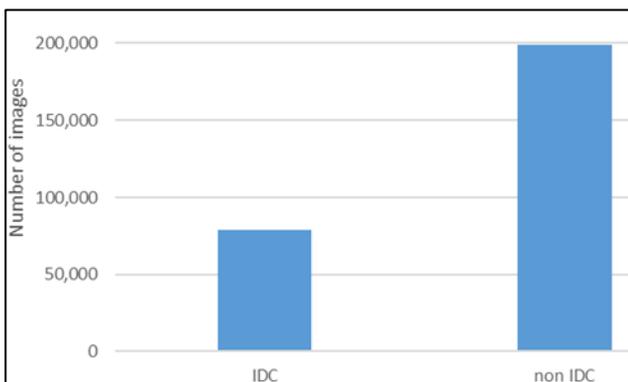


Fig. 3. The dataset distribution.

In this proposed method, to load and manipulate the images, the library image in Keras package is used. Then simple preprocessing is applied to stack all 50x50 images into 4D array to be able to deal with it in the implementation. Next,

the pixel values are normalized. The reason is that the pixel values can range from 0 to 256, where each number indicates a gray level value. The computation of large numeric values may get more difficult when sending these values through CNNs. We may lessen this by normalizing the numbers to a range of 0 to 1 by dividing the array by 255.

5) *Splitting dataset.* Train-Test split is a technique to evaluate the performance of the proposed model with giving 20% for the test set. The method *train-test-split()* in sklearn helps to split the images into training and test sets. The training set is used to train SVM classifier and then it calculates the accuracy of the training set as an average over 10-cross-validation. The training is also used to draw a learning curve. The test set is used to test the trained SVM classifier and then calculates the accuracy of the test set as an average over 10-cross-validation.

### C. Results

Different experiments were concurred to investigate the pretrained models and analyze the results trying to get a better performance. The best results are written down in this paper.

Table VII-A shows the overall comparison of the accuracy of the three cases. It shows the accuracy of training and test sets over the three classifiers from the three models. The values represent the mean of the accuracy which is the average value of accuracy among 10 different sizes of the training set along with their standard deviation values to represent the dispersion of accuracy values around the mean. The Table VII-B and Table VII-C represent values in the same manner as Table VII-A but investigate different effects. Table VII-B investigates the effect of swapping between the two steps of the methodology, Step 3 and Step 4. While Table VII-C investigates the effect of change the up-sampling methods. Table VIII shows the effect of applying standardization to the features on the accuracy of the combined SVM classifier, where the time is in seconds.

Fig. 4 shows the learning curve of the training and validation accuracy of the trained SVM classifier for varying numbers of training images. The x-axis shows the number of images that will be used to generate the learning curve. The y-axis shows the average of the accuracy values over 10 runs for each training subset size. The training and validation accuracies for different training set sizes in 10-cross-validation is measured to investigate influence of number of images on accuracy of the SVM classifier. Recall that the SVM classifier in this case is the combined SVM classifier.

### D. Discussion

Table VII-A shows both VGG16 and ResNet50 models give a satisfactory performance when using a feature hierarchy. The initialed experiments started without hierarchy, i.e., using only the last convolution layer as features but the result was lower. The result corresponds to the research [26], which confirms that the concept of feature hierarchy can be successfully applied to breast cancer classification. At the same time, the performance is better when the features from the two models are fused. This result shows the effect fusing of different pretrained models to get a better result than using each pretrained individually.

TABLE VII. THE ACCURACY OF THE THREE CLASSIFIERS FROM THE THREE MODELS

	Data sets	VGG16	ResNet50	Combination
A) Original	Training set	0.927±0.015	0.923±0.016	0.943±0.014
	Testing set	0.897±0.028	0.889±0.031	0.920±0.041
B) After swapping Step 3 & 4	Training set	0.923±0.013	0.923±0.015	0.943±0.021
	Testing set	0.892±0.030	0.890±0.036	0.908±0.033
C) After double operation	Training set	0.923±0.018	0.926±0.017	0.933±0.019
	Testing set	0.899±0.025	0.899±0.029	0.916±0.039

TABLE VIII. THE ACCURACY OF THE TWO CLASSIFIERS WITH/WITHOUT APPLYING THE STANDARDIZATION

Applied?	Data sets	Accuracy	Fit time	Fold time	Warning?
Yes	Training set	0.943±0.014	144.05 s	1054.24 s	No
	Testing set	0.920±0.041	-	128.61 s	No
No	Training set	0.897±0.011	186.02 s	1300.88 s	Yes
	Testing set	0.915±0.023	-	126.19 s	No

According to pretrained models, we can also observe from Table VII-A that the activations of pretrained model, that are trained on ImageNet dataset [32], can be used as features in the proposed model in the breast classification task with consuming less implementation cost, i.e., using pretrained models instead of training from scratch.

Moreover, the extracting steps shown on Fig. 2 describe applying up/down sampling (Step 3) before applying depth pooling (Step 4). However, both these two steps are related to unifying the shape of feature maps. Thus, if these two steps are swapped. i.e., applying up/down sampling after depth pooling, the result is almost the same with a small enhancement for the original case as shown on Table VII-B. One of the possible reasons is both operations work on getting the maximum value which will generate almost similar values in two directions.

Regarding up-sampling operations, different operations can be used other than the transpose convolutional layer. The other simple common type is using the method *UpSampling2D()* to double the dimensions of the input. After applying the simple double operation in the experiments, it gives almost the same result as Table VII-C shows with little enhancement for the original case. However, the key difference is in their learning. The simple double operation is a simple scaling up of the input without learning to achieve a less complicated in implementation. Whereas the transpose convolutional operation is a convolution operation whose kernel is learnt while learned the model to learn the best up-sampling for the task.

Turning to the normalization on Table VIII, the case of standardization (the top case) gives a faster result in fewer iterations. It can be considered as one of the solutions to overcome the warning of convergence issue that expresses that the estimation terminated early before reaching the convergence. Also, standardization achieves a better result in terms of accuracy. Especially for SVM classifier, the scaling helps to decrease the distances between inputs to select the maximum decision boundary.

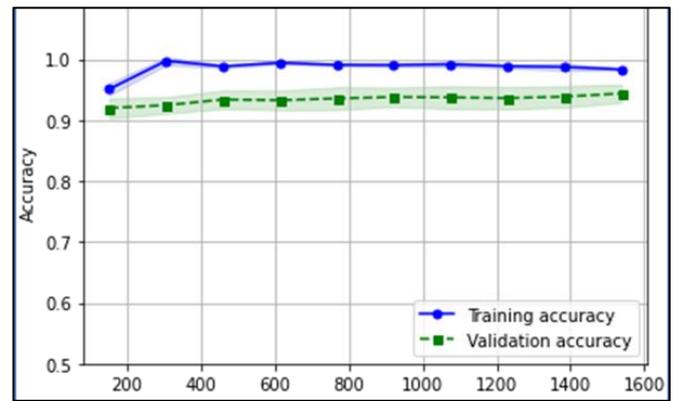


Fig. 4. The evaluation of the classifier over different examples.

The overall trend of Fig. 4 shows the effect of the number of images on the accuracy. The accuracy of the training set is higher than the validation set but with acceptable gap, i.e., the gap between them did not increase after a specific point to express about happening overfitting. One of the possible reasons is using the pretrained models with their learnable parameters, i.e. all layers are frozen and used the same weights. This helps to reduce the number of required images to reach the convergence.

Some other notes are appeared during the experiments. There are different possible sizes of the receptive field can be chosen to unify the shape of feature maps. When trying to unify the size to be 3x3, the resulting accuracy is almost the same in most cases, with little increase for 6x6 in other cases. However, more investigations in the future are better to be conducted on this size and other varied sizes to get adequate results about the effect of changing size on performance.

#### E. Comparing with the State-of-the-art

Ultimately, the results correspond to the research [26], which observes a good effect of using a feature hierarchy. While the proposed model in this paper uses the feature hierarchy for different domain, which is the breast cancer classification tasks.

At the same time, comparing to the other works provided in Section II, this proposed model seems to infer the same previous result [6] [7] about the effect fusing of different pretrained models to get a better result than using each pretrained individually. But in this paper, the combination is between a features hierarchy extracted from two models VGG16 and ResNet50.

#### V. CONCLUSION

IDC is the most common subtype of all breast cancers. Instead of manual diagnosis, it must find solutions to ease diagnostic burdens, especially in under-staffed laboratories and equipment. Thus, the goal of this paper is to classify breast cancer in the form of images into binary classification: IDC and not IDC. This paper proposes a CNN-based model for learning features of breast cancer images that combines two pretrained CNN models to extract a feature hierarchy and then feeds them into the SVM classifier. Besides, experimental results show that classification performance is higher in the

combined pretrained model and fusing the deep features from various layers from various pre-trained CNNs leads to better classification performance. In addition, other findings present the effect of some factors such as the normalization of training SVM classifier. However, those results are not the best results. It can be considered as a contribution, while the performance can be after the additional investigation in several factors, such as change the size of the receptive field of the features maps, number of pretrained models as well as other datasets with different pixel sizes may get another improvement.

In the future, this paper provides various recommendations that are expected to help in developing CNN models. First, combining other information along with the breast images during developing DNN models, such as changes in the breast shape and DNA sequences, may increase the accuracy of the classification. Second, the breast imaging modalities are better to consider during developing DNN models. Adopting new modalities of imaging may provide more accurate details, such as shear wave elastography (SWE) or magnetic resonance imaging (MRI). Third, while the enormous quantity of unlabeled photos is a valuable source of data, it cannot be used in supervised learning. Instead, the research can shift to training in an unsupervised manner, such as using clustering approaches. In the end, increasing research interest and rapid technological advancements creates a chance for researchers to continue to evolve models of breast cancer classification.

#### REFERENCES

- [1] Nahid and Y. Kong. "Involvement of machine learning for breast cancer image classification: a survey." Computational and mathematical methods in medicine 2017, 2017.
- [2] B Sahiner, H Chan, N Petrick, D Wei and M. Helvie. "Classification of mass and normal breast tissue: a convolution neural network classifier with spatial domain and texture images". IEEE Trans Med Imaging, vol. 15, no. 5, pp. 598-610, 1996.
- [3] K. Simonyan and A. Zisserman. "Very deep convolutional networks for large-scale image recognition." arXiv preprint arXiv:1409.1556, 2014.
- [4] H. Kaiming, 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, pp. 770-778. 2016.
- [5] C. Cortes, and V. Vapnik. "Support-vector networks." Machine learning, vol. 20, no. 3, pp. 273-297, 1995.
- [6] S. Khan, N. Islam, Z. Jan, I. Ud Din, and J. Rodrigues. "A novel deep learning based framework for the detection and classification of breast cancer using transfer learning." Pattern Recognition Letters 125, pp: 1-6, 2019.
- [7] A. Mahbod, G. Schaefer, C. Wang, R. Ecker, and I. Elling. "Skin lesion classification using hybrid deep neural networks." In ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1229-1233. IEEE, 2019.
- [8] R. Mehra, "Breast cancer histology images classification: Training from scratch or transfer learning?". ICT Express, vol. 4, no. 4, pp. 247-254, 2018.
- [9] R. Andrik, W. Scotney, J. Morrow, and H. Wang. "Breast mass classification in mammograms using ensemble convolutional neural networks." In 2018 IEEE 20th International Conference on e-Health Networking, Applications and Services, pp. 1-6. IEEE, 2018.
- [10] S. Shamy, J. Dheeba, "A research on detection and classification of breast cancer using k-means GMM & CNN algorithms," International Journal of Engineering and Advanced Technology, vol. 8, no. 6S, pp. 501-505, 2019.
- [11] P. Lloyd "Least squares quantization in PCM", IEEE Transactions on Information Theory, vol. 28, no. 2, pp. 129-137, 1982.
- [12] V Sansya Vijayan, Lekshmy P L, "Deep learning based prediction of breast cancer in histopathological images," International Journal of Engineering Research & Technology, vol. 8, no. 07, pp.148-152, 2019.
- [13] M. Rathi and V. Pareek, "Hybrid approach to predict breast cancer using machine learning techniques," International Journal of Computer Science Engineering, vol. 5, no. 3, pp. 125-136, 2016.
- [14] S. Christian, W Zaremba, I. Sutskever, J. Bruna, D. Erhan, Ian Goodfellow, and R. Fergus. "Intriguing properties of neural networks." , arXiv preprint arXiv:1312.6199, 2013.
- [15] A. Krizhevsky, I. Sutskever and G. Hinton, "ImageNet classification with deep convolutional neural networks", Communications of the ACM, vol. 60, no. 6, pp. 84-90, 2017.
- [16] K. Simonyan, A. Zisserman, "Very deep convolutional networks for largescale image recognition", in: International Conference on Learning Representations, pp. 1-14, 2015.
- [17] Lloyd, S. P."Least square quantization in PCM". Bell Telephone Laboratories Paper. Published in journal much later: Lloyd., S. P, 1982.
- [18] D. E. Goldberg. "Genetic Algorithms in Search, Optimization and Machine Learning". New York: Addison Wesley, 1989.
- [19] Ding C, Peng H. Minimum redundancy feature selection from microarray gene expression data. J Bioinforma Comput Biol, vol. 03, no. 02, pp. 185-205, 2005.
- [20] Lo C-S, Wang C-M. "Support vector machine for breast MR image classification". Comput Math Appl, vol. 64, pp :1153-1162, 2012.
- [21] J. Ren "ANN vs. SVM: which one performs better in classification of MCCs in mammogram imaging", Knowl-Based Syst, vol. 26, pp :144-153, 2012.
- [22] M. Tahmoorei, A. Afshar, B. Bashari Rad, K. B. Nowshath, M. A. Bamiah, "Early detection of breast cancer using machine learning techniques," Journal of Telecommunication, Electronic and Computer Engineering, vol. 10, no. 3-2, pp. 21-27, 2018.
- [23] Ebru Aydindag Bayrak, Pinar Kirci, TolgaEnsari,"Comparison of machine learning methods for breast cancer diagnosis". 2019 Scientific Meeting on Electrical-Electronics & Biomedical Engineering and Computer Science (EBBT), pp. 1-3,2019.
- [24] Ch. Shravya, K. Pravalika, ShaikSubhani, "Prediction of breast cancer using supervised machine learning techniques," International Journal of Innovative Technology and Exploring Engineering, vol. 8, no. 6, pp. 1106-1110, 2019.
- [25] C. Aroef, Y. Rivan, and Z. Rustam. "Comparing random forest and support vector machines for breast cancer classification." Telkomnika 18, no. 2, pp. 815-821, 2020.
- [26] C. Ma, J. Huang, X. Yang and M. Yang, "Robust Visual Tracking via Hierarchical Convolutional Features," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 41, no. 11, pp. 2709-2723, 2019.
- [27] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, and M. Bernstein, "Imagenet large scale visual recognition challenge," International Journal of Computer Vision, vol. 115, no. 3, pp. 211-252, 2015.
- [28] B. Harangi, "Skin lesion classification with ensembles of deep convolutional neural networks", Journal of Biomedical Informatics, vol. 86, pp. 25-32, 2018.
- [29] L. Liu, C. Shen, & A. van den Hengel. "Cross-convolutional-layer pooling for image recognition". IEEE transactions on pattern analysis and machine intelligence, vol. 39, no. 11, pp. 2305-2313, 2016.
- [30] Hastie, T., Tibshirani, R. and Friedman, J., The elements of statistical learning. 2nd ed. New York: springer. 2017.
- [31] Koo, KM., Cha, EY. "Image recognition performance enhancements using image normalization". Hum. Cent. Comput. Inf. Sci. vol. 7, 2017.
- [32] "About ImageNet", image-net, 2022. [Online]. Available: <https://www.image-net.org>. [Accessed: 01- Mar- 2022].
- [33] "Breast Histopathology Images", Kaggle, 2022. [Online]. Available: <https://www.kaggle.com/datasets/paultimothymooney/breast-histopathology-images?resource=download>. [Accessed: 01- Mar- 2022].

# A Survey on Attention-Based Models for Image Captioning

Asmaa A. E. Osman<sup>1</sup>, Mohamed A. Wahby Shalaby<sup>2</sup>, Mona M. Soliman<sup>3</sup>, Khaled M. Elsayed<sup>4</sup>

Information Technology Department-Faculty of Computers and Artificial Intelligence,  
Cairo University, Giza, Egypt<sup>1,2,3,4</sup>

Smart Engineering Systems Research Center (SESC), Nile University, Giza, Egypt<sup>2</sup>

**Abstract**—Image captioning task is highly used in many real-world applications. The captioning task is concerned with understanding the image using computer vision methods. Then, natural language processing methods are used to produce a description for the image. Different approaches were proposed to solve this task, and deep learning attention-based models have been proven to be the state-of-the-art. A survey on attention-based models for image captioning is presented in this paper including new categories that were not included in other survey papers. The attention-based approaches are classified into four main categories, further classified into subcategories. All categories and subcategories of the attention-based approaches are discussed in detail. Furthermore, the state-of-the-art approaches are compared and the accuracy improvements are stated especially in the transformer-based models, and a summary of the benchmark datasets and the main performance metrics is presented.

**Keywords**—Image captioning; attention model; deep learning; computer vision; natural language processing

## I. INTRODUCTION

Image captioning is targeted to represent an image with a sentence that should be accurate and summarized. The problem of image captioning is similar to using a machine to translate a sentence, but in image captioning, the machine task will be translating an image into a sentence. So, it is necessary to visually understand the image before producing the caption. The caption of the image should be expressive through detecting the objects of the image and their attributes, finding the relationship between the detected objects and the place/activity where the objects are included.

The task of image captioning is very necessary for that it can be as an assistant to the impaired people by providing a brief description for the image while exploring the internet. Image captioning can be used in implementing self-driving cars by providing the agent with the ability to drive in a safe, fast and accurate way. Also, generating a caption for medical images automated the process of diseases diagnosis and treatment. In addition, it can be used to generate captions for the images included in the news articles. There are many other applications for image captioning, like in service robotics, military, education and image indexing.

In order to generate a sentence with reasonable linguistics and true semantics, Computer Vision (CV) methods are used to visually understand the image. In addition, Natural Language Processing (NLP) models are employed to generate a correct

sentence. The power of Deep Learning (DL) approaches in CV [1-7] and NLP [8-12] makes it the first choice for many approaches in image captioning. Convolutional Neural Network (CNN) was most commonly used in the vision part to get the image features. Then, Recurrent Neural Network (RNN) was used as a language model [13-16].

According to [17], deep neural network approaches in image captioning task can be categorized based on:

- Type of learning: (Supervised [18,19], Unsupervised [20,21] and Reinforcement Learning [22,23])
- Architecture: (Encoder-Decoder [24,25] and Compositional [26,27])
- Feature Mapping: (Visual Space [28,29] and Multimodal Space [30])
- Number of Captions: (Dense Captioning [31], Whole Scene Captioning [32])
- Language Model: (LSTM and others)

For the purpose of generating high quality captions, it was helpful to use advanced visual processing by considering the most salient features in the images while generating the caption words which is called attention model. The attention mechanism takes inspiration from the human visual system, which does not focus on all the scene parts but only on small parts of the scene. The salient features included in the image take precedence in encoding the image instead of the whole image. Attention has been used in different tasks, like machine translation and object identification. Moreover, many image captioning approaches employed the attention model and achieved a very good enhancement [33-37].

In this paper, a detailed survey for the attention-based approaches employed in image captioning is presented. In addition, a taxonomy of these attention-based models is provided including two new categories for categorizing the attention-based approaches. Most of the state-of-the-art articles for image captioning using attention-based models are included and compared with respect to the benchmark datasets and metrics.

The organization of this survey paper is as follow: In Section II, Literature review is presented. The attention mechanism and its taxonomy is presented in Section III, including four main categories of the attention models and their subcategories. The benchmark datasets in addition to the

popular performance metrics are introduced in Section IV. State-of-the-art models are compared in Section V. Finally, this survey paper is concluded in Section VI.

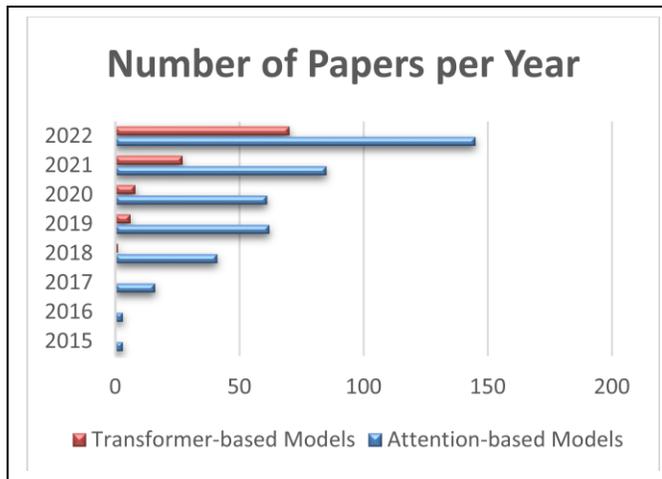


Fig. 1. Number of attention-based papers in image captioning per year

## II. LITERATURE REVIEW

Many surveys have been published for the deep learning techniques in image captioning [17, 38-42]. Some of these surveys [17, 38, 40-42] considered few attention-based approaches in image captioning because most of the attention approaches were issued after publishing these deep learning surveys. A comparative study for attention-based techniques was published by Khaing and Phyu [43]. Their survey presented a good comparative study of the attention-based models but without any categorization and moreover, the most recent reference in their survey was in the year 2018, and there is big progress in the attention-based methods starting from the year 2019 as shown in Fig. 1. The newest survey for attention-based models was presented by Zohourianshahzadi and Kalita [44]. In [44], they presented an evolution path of the attention models including hard and soft attention, semantic attention, spatial attention, adaptive attention, and bottom-up and top-down attention.

As per our knowledge, there is no detailed survey with a good taxonomy for the attention-based approaches employed in image captioning. Motivated by this gap in the existing image captioning survey papers, especially for the attention-based approaches, a detailed survey for the attention-based approaches employed in image captioning is presented in this paper by introducing new categories.

## III. TAXONOMY OF ATTENTION-BASED MODELS

Employing the attention mechanism in image captioning was motivated by the successful work achieved in neural machine translation [45] and object recognition [46, 47]. The attention was employed in the decoder part of the translation task to mitigate the encoder from the need to model all input sentence information [45]. Xu et al. [48] proposed captioning

approach by exploring the attention technique to consider the significant regions in the process of caption generation.

According to [48], the attention was applied at the decoder so that at every time step ( $t$ ), LSTM produced a new word depending on the hidden state ( $h_{t-1}$ ), the words produced at the previous steps and a vector called context vector ( $\hat{z}_t$ ). The context vector ( $\hat{z}_t$ ) represents the information of an appropriate location of the image at specific time step  $t$ . The context vector  $\hat{z}_t$  can be calculated using the annotation vectors, which are the features related to the image regions, and their assigned weights  $\alpha$ . The weights  $\alpha$  are assigned to every annotation vector  $a_i$ , ( $i = 1, \dots, L$ ) using Multilayer Perceptron depending on the previous step hidden state  $h_{t-1}$ . The attention model  $f_{att}$  used for calculating the weights had two variants either soft or hard attention depending on how the weights will be interpreted.

Variants of the attention model were proposed in image captioning research area, some researchers enhanced the model by employing the attention as multi-stages or by inserting information to guide the attention. However, the most notable variant of the attention is the transformer-based models as can be seen from Fig. 1, there is a big interest in applying the transformer-based models in comparison with the other categories.

In image captioning, the attention mechanisms can be categorized into four categories, as demonstrated in Fig. 2. According to Chen et al. [49], the visual attention-based approaches may concentrate on the spatial features or the semantic features, so visual attention is added as a category for characterizing the attention-based models. In addition, according to He et al. [50], the attention-based methods can be categorized based on applying the attention as single-stage in the decoder, two-stages, two-stages with scene graph or based on the transformer. This classification is added as subcategories into the category named Attention Blocks. In addition to these main two categories, in this survey paper, two new categories that were not included in other survey papers for characterizing the attention-based models are added, which are Number of Attention Layers and Guided-Attention.

### A. Visual Attention

Visual Attention [51, 52] is a significant technique in the human visual system. The brain targets a region or an object using computational capabilities with the guidance of low-level image features in a time step. The visual attention models can be divided into spatial and semantic attention.

1) *Spatial attention*: For spatial attention, the attention is demonstrated spatially at a specific region [48, 49, 53-56]. For each fixed location, attention weights are calculated related to this location at each iteration. Several approaches apply soft attention, which models the feature maps with the computed weights.

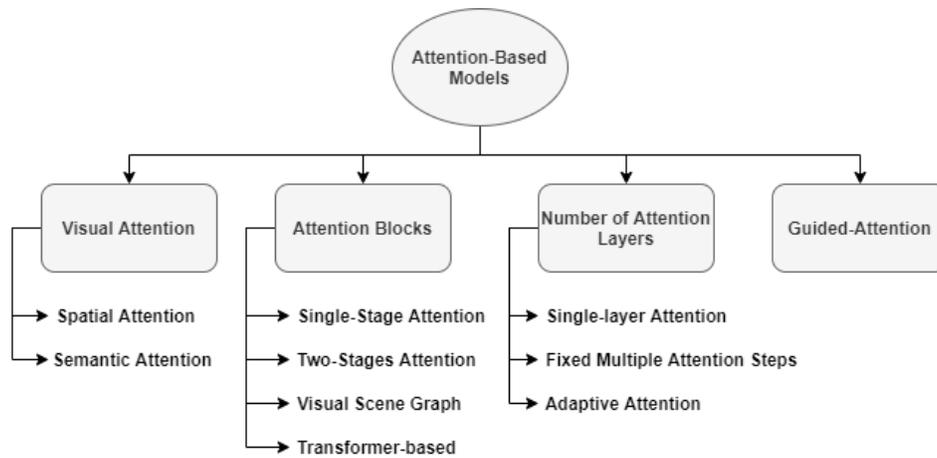


Fig. 2. Taxonomy of attention-based image captioning models

While other approaches use hard attention by selecting a set of regions which are salient from the feature map and concealing the other regions. Through applying the weighted pooling, some of the important spatial data may be lost. In addition, regularly the spatial attention is computed in the last convolutional layer, which leads to some analogous feature results for distinct regions because of the big size of the filter, resulting in ineffective spatial attention.

2) *Semantic attention*: Instead of attending to the fixed resolutions in spatial attention, other approaches proposed attending to the image's semantic concepts [57, 58]. Semantic attention is more like the human description of the image because people describe the most important objects and do not talk about all regions in the image. Attributes can be utilized from any image location even if there is no actual existence of these attributes within the image. For the purpose of attending to semantically necessary attributes, You et al. [57] employed a semantic attention framework that used top-down and bottom-up models. Bottom-up was used to select the semantic attributes, and top-down was used to decide when and where to apply the attention. Another approach was proposed by Gan et al. [58] in which they recognized the semantic tags and computed the probability of the tags to be utilized in forming the LSTM parameters. LSTM weight matrices were expanded to a group of weight matrices that are tag-dependent.

Using semantic attention requires extra resources that are important for detecting the relationship among the semantic concepts and the image.

### B. Attention Blocks

The attention-based models can also be categorized according to the block where the attention is applied. The attention can be applied as a single-stage in the decoder block, two-stages by obtaining bottom-up and top-down attentions, two-stages with injecting a graph network, or Transformer-based models.

1) *Decoder-based attention (single-stage attention)*: In decoder-based attention models, the attention is employed at the decoder. In the process of producing the caption words, the informative regions [59] are targeted in the attention by the

decoder. Depending on the LSTM hidden states and the previously predicted caption words, Xu et al. [48] proposed to use the attention module in the decoder of the captioning approach while generating the sentence words. A weighting matrix is introduced for each feature map receptive field then this weighted map and the last predicted word were forwarded to the language model for the purpose of predicting the next word.

2) *Two-stage attention*: Rather than attending to the salient regions like Decoder-based attention, Anderson et al. [53] presented a model that contains two-stage attention. Faster R-CNN [60] was employed in the bottom-up attention module. Then, the attention was distributed among the image regions using a top-down attention mechanism. They used two LSTM layers for the purpose of applying attention to the selected spatial features, the first layer was for the top-down attention, and the other was for the language layer. The drawback of their model is that it cannot handle object-object relationships.

An approach that is similar to [53] was introduced by Lu et al. [61]. Their proposed decoder determines whether the word will be visual and predicted according to a certain image region or the word will be predicted from the textual vocabulary. The essential advantage of their approach is in its availability to have additional object detectors, which can lead to producing different image captions. The main gap in two-stage attention models is that the models are lacking for getting the relationship between the image regions.

3) *Two-stage attention with graph*: To enhance the two-stage attention models, graph networks can be employed to discover the relationship among the detected regions which can result in enhanced features and accordingly improve the caption generation. Similar to [53, 62], Yao et al. [55] employed the attention mechanism for attending to the informative image regions. The key novelty in (GCN-LSTM) [55] is that they used two graphs for detecting the relationship between the image regions. A semantic graph was employed with the nodes representing the image regions and the edges representing the relationship between these detected image

regions. While the geometrical relations between the regions 'vertices' were demonstrated by the spatial graph. Then, Graph Convolutional Network (GCN) [63] was utilized to output relation-aware region representations.

The approaches presented in [55, 64] employed Faster R-CNN to identify the image objects and thus explore the relationships between regions of interest. Faster R-CNN was trained on the Visual Genome dataset [65]. While in [66], the visual relationships were modelled on Flickr30K [67] and MS COCO [68] and so the pre-established classes of the relations are not required.

The authors in [55] extended the approach to (GCN-LSTM-HIP) [69] to include a hierarchical tree of three levels which have the image as the root, the detected regions as the first layer and the instances/foreground of the regions at the leaf layer. Then, a Tree-LSTM [70] was employed for modelling the dependency structure and improving the features.

Another model presented by Guo et al. [71] which detected a set of visual semantic units 'VSUs' where the units represent the objects, attributes and the object's relationships. Semantic and geometry graphs were employed while the vertices representing the semantic units and the edges representing the connections between them differed from [55] that presented the relationships as edges. GCN was then introduced in [71] to output context-aware embeddings for the visual semantic units. Attention for the different kinds of units was applied via context gated attention (CGA). Another scene graph approach was presented by Yang et al. [72] that used the edges to represent the relationships in the graph. Language inductive bias was integrated into the captioning framework, and its features are represented via a scene graph auto-encoder (SGAE).

The main drawback in the graph scene-based models is that, however, the models made an enhancement to the performance compared to the two-stage models, but the need for additional models for scene graph construction is still a problem. Also, with respect to the computational cost, having two graphs is ineffective.

4) *Transformer-based*: Unlike the graph-based models, the transformer models don't include any graphs and thus don't need additional models for the graph construction. The transformer was originally designed for text translation [73]. The transformer is able to avoid any duplication by employing the attention in a comprehensive way between the input and the output. Extensive approaches were proposed to employ the transformer models in image captioning [74-85].

Huang et al. [86] proposed Attention over Attention (AoA) approach, which adds attention over the traditional attention. "Information vector" and "attention gate" were produced by the query and the attended results, then second attention was produced by element-wise multiplication between them. AoA was applied in the encoder to detect the relations between the objects. While in the decoder, AoA was employed for holding the relevant attention output and ignoring the deceptive results.

Captioning transformer with stacked attention module was proposed by Zhu et al. [76]. A multi-level observation was proposed in such a way that all transformer layers had the

opportunity for generating the sentence word. Average pooling was then employed to find the probability of the word by merging all the contributions.

Cornia et al. [74] proposed a transformer approach to consider low and high-level relationships by modelling them as multi-level. They utilized persistent memory vectors while encoding the relationships with prior information. In addition, rather than applying the attention only to the last encoding layer, all the encoder layers contributed to the sentence generation process and connected to the decoder layers in mesh-like connectivity.

A Multimodal transformer was proposed by Yu et al. [75], which is able to model three different relations, which are: word-to-word, word-to-object and object-to-object. Self-attention in the same modality and co-attention in distinct modalities were acquired. In addition, multiple views were employed in two designs: aligned and unaligned multiple views.

The conventional transformer was expanded with the addition of EnTangled Attention (ETA) and Gated Bilateral Controller (GBC) [77]. ETA gave the transformer the ability to use semantic concepts and visual information. The interconnection between the multimodal information was controlled by the GBC. Object relation transformer [78] was proposed in which geometrical information for the relationship between each pair of objects was included within the transformer through spatial attention.

He et al. [50] proposed a model with the idea of changing the internal structure of the transformer that was originally proposed to handle text. They introduced an expanded transformer that includes three parallel sub-transformer layers to handle three different relationships: parent, child, and neighbor.

### C. Number of Attention Layers

The attention models can be characterized according to the number of required attention steps either to attend once per word, attend with fixed steps or adaptively determine the number of required attention steps.

1) *Single-layer attention*: The attention operation is connected to the word generation procedure in the traditional attention-based framework [48]. The framework attended once to the image prior to generating the following word. The model attended to selected image regions in each iteration, and the computed attention features were sent to the RNN as input. The problem of attending once per word is that some important information may be lost, especially if the model attended to an incorrect region.

2) *Fixed multiple attention steps*: In order to enhance the single attention process by avoiding predicting incorrect words, several approaches attended multiple times to enhance the attended region and get the lost data [87, 88]. Du et al. [87] proposed a model that attended more times to the image per word and showed that it could improve image captioning without adding extra parameters. Two LSTMs model were

utilized, which have the ability to attend for arbitrary times and enable the flexibility of the attention operation.

Triple attention approach was proposed by Zhu et al. [88]. The attention is utilized to the input phase of the previous step LSTM hidden states. In addition, attention was also utilized in the output phase of present hidden states. Conditional embedding was used in addition to the word/image embedding at every input stage of LSTM. This way, the prior text information was coupled with image information, and accordingly, text and image information appeared in the input of the word generation procedure.

For the purpose of getting attention to different semantic abstractions, Chen et al. [49] applied the attention in a multi-layer since the lower layers are the dependent layers for the feature maps. The attention in their approach was approached to each entry of the feature maps, which are multi-layer. They also proposed channel-wise attention for applying the re-weighting process in every channel through the word generation process. The channel-wise attention could be viewed as the procedure of choosing the semantic concepts by paying more attention to the channels produced by filters indicated by the semantics.

A hierarchical approach (CNN+CNN) [89] was proposed such that they employed the CNN as a decoder besides being the encoder. Their hierarchical attention model learns the relationship of the attributes for all image regions and all levels. The dot-product operation used in their framework results in reducing the parameters and can be faster than Multi-layer Perceptron attention used in [48, 54]. The idea of hierarchical attention was also employed in [90-92]. Yan et al. [90] proposed a mechanism made up of global and local attention modules which related to the global CNN features, extracted by CNN encoder, and local object features, extracted by object detector, respectively.

Sequential attention was presented by Fang et al. [93] to take into consideration the sequential attention relationships in several time steps at word generation and correspondingly improve the visual data in caption generation. Another sequential attention was proposed by Liu et al. [94], in which the image was represented as a sequence of objects, and the attention was employed to consider all objects information during sentence word generation.

3) *Adaptive attention*: According to the previously presented approaches, sometimes there are no image regions corresponding to each sentence word. So, an adaptive attention approach [54] was proposed that includes a sentinel gate and spatial attention to determine where and when to attend in the caption generation. They presented an extension to LSTM that, rather than having one hidden state, they added a visual sentinel vector. In addition, a sentinel gate was proposed to determine whether the attention will be targeted to the visible sentinel or the image. Another adaptive attention approach was proposed by Deng et al. [95] that adaptively determine whether it is needed to depend on the language model or the visual signals. Their proposed approach can make the image captioning task more flexible by enhancing

the obligatory correlation between image regions and sentence words.

Adaptive semantic attention framework [96] was proposed to incorporate dual-LSTMs; the first LSTM works as a visual sentinel to acquire fine-grained representations. The second LSTM serves as a language model that produces the sentence words depending on the updated attended vector and first LSTM output.

Huang et al. [97] presented an adaptive attention time model (AAT). The model was learned to determine the number of required attention steps in each step of the decoder in order to produce the next word. Using AAT, the mapping between the image regions and caption words can be applied arbitrarily such that a caption word may attend to multiple regions and vice versa. Their approach doesn't add parameters gradient noise.

#### D. Guided Attention

For the purpose of enhancing the performance of image captioning approaches and generating accurate captions, some approaches inserted additional information guidance [98, 99] like the concept features that make a connection between the input image and the caption. In [100], the model was guided through semantic information acquired from the images and sent as extra input for the LSTM units. While in [101], the approach could be guided through concept features which are obtained from predicting the recurrent word existence in the captions. Another way for learning the features is by adding a network for guidance [102]. More similar to [102], Sow et al. [103] inserted a network for guidance, but rather than obtaining one vector for guidance, [103] obtained a sequential network for guidance which was able to adjust the guided vectors in the sentence generation process. They also utilized the Luong attention mechanism [104] that is an enhanced style of the attention technique.

Text-guided attention approach was presented by Mun et al. [105]. Related sample captions, namely guidance captions, were employed to get visual attention and produce appropriate captions. The related sample captions were obtained through the similar training images that participate in equivalent related regions with the input image. Topic-guided attention was proposed by Zhu et al. [106], which picked up the significant features by the information guidance through incorporating the topics within the image with the attention mechanism.

## IV. DATASETS AND PERFORMANCE METRICS

### A. Datasets

Different datasets have been presented in the research area of image captioning. The popular datasets, which are Flickr8K [107], Flickr30k [67], Microsoft COCO [68] and Visual Genome [65] are presented.

1) *Flickr8K [107]*: Dataset consists of about eight thousand images selected from six groups on Flickr.com and does not have a tendency to famous locations or people; instead, various situations and locations are represented. The dataset includes five captions for each image through human annotations.

2) *Flickr30K* [67]: Extension to Flickr8K, consists of 31,783 images. Flickr30k contains 8.7 objects per image, 44,518 object categories, 6.2 objects per category, 5 sentences per image and 16.6 expressions per image.

3) *Microsoft COCO dataset* [68]: A large-scale dataset that broadly used in image captioning task. MS COCO includes 328,000 images, 7.7 objects per image, 91 object categories, 2.5 million labelled instances, 27,473 objects per category and five sentences per image.

4) *Visual genome dataset* [65]: It is an image captioning dataset that considers the relationship modelling between objects. It generates captions for different image regions, unlike the other datasets, which generate the caption to the entire scene. The dataset includes more than 100 thousand images, 18 attributes, 21 objects per image, and 18 objects relationships.

### B. Performance Metrics

For the purpose of evaluating the image captioning techniques, different metrics were proposed to compare the output generated caption with the original caption. In this section, the main used performance metrics which are BLEU [108], ROUGE [109], METEOR [110], CIDEr [111] and SPICE [112] are presented.

1) *BLEU "Bilingual Evaluation Understudy"* [108]: It is originally introduced by IBM for the evaluation of machine translation. This metric measures the quality of the generated sentence by calculating its similarity with the original reference translations. N-grams of the machine-generated sentence are compared to those of the reference sentences and get the matching counter. The output score is higher, and the quality of the generated sentence is better when there are more reference sentences and there is a higher number of matches. The range of BLEU values is from zero to one, and a small number of generated captions can get one only if it is identical to the ground truth caption.

2) *ROUGE* [109]: It is originally introduced for the evaluation of text summarization. ROUGE metric calculates the quality of the text generated summary by counting the number of its n-gram, sequences of words, and pairs of words that overlapped with the reference summaries created by experts. ROUGE-N (N-grams), ROUGE-L, ROUGE-W, and ROUGE-S are the types of the ROUGE metric.

3) *METEOR* [110]: A metric utilized for evaluating the machine-generated texts by matching the unigrams of the machine-generated sentence and the reference sentences. Once this matching is computed, recall and precision of unigram and a measure of fragmentation were utilized for computing a METEOR score.

4) *CIDEr* [111]: A metric utilized for evaluating the image descriptions. The five available captions of the dataset used in the other metrics are not enough for finding the consensus among the judgment of the human and the output captions. A consensus is a measurement for counting the

mutual n-grams between the ground truth and predicted captions and assigning low weights for the common n-grams.

5) *SPICE* [112]: The previously explained metrics depend on the n-grams and SPICE metric overcomes this restriction by employing a scene graph in which the reference and generated captions are converted to a graph-based semantic representation. SPICE is measuring if the objects and attributes are represented in the generated caption in an effective way in addition to their relationships.

## V. COMPARISON AND DISCUSSION

In this section, the performance of different state-of-the-art approaches is presented and discussed. In Table I, different approaches are compared with respect to their experimental results on the benchmark MS COCO dataset and the commonly used performance metrics BLEU-4 (B@4), METEOR (MT), ROUGE-L (R), and CIDEr (C).

From the beginning of using the attention mechanism in image captioning by Xu et al. [48], it has been shown that their approach obtained better performance on Flickr8k, Flickr30k and MS-COCO. The reason behind the better performance is that their approach considered the most relevant objects when generating the image caption. Moreover, they showed that the hard attention variant of their mechanism outperforms the soft attention on these benchmark datasets. After that, You et al. [57] showed that attending to the semantic attributes instead of attending to the spatial attention [48] can improve the results by generating semantically rich captions.

Further improvement in the results was obtained by introducing multiple attention layers, which can be used in a hierarchical structure or by using either a fixed or adaptive number of attention layers. Du et al. [87] achieved 38.1, 28.3, 58.0, 126.1 and 22.0 on BLEU-4, METEOR, ROUGE, CIDEr and SPICE, respectively. These results are higher than the results of hierarchical attention [89]. The hierarchical structure in [89] used the CNN as decoder; however, Du et al. [87] used two LSTMs model to enable attention at arbitrary times and make the attention operation more flexible.

The adaptive attention approach of Huang et al. [97] achieved 38.7, 28.6, 58.5, 128.6 and 22.2 on BLEU-4, METEOR, ROUGE, CIDEr and SPICE, respectively, which are higher than that of both Wang and Chan [89] and Du et al. [87]. The reason for their better performance is that their model was learned to determine the number of required attention steps in each decoder step, and the mapping between the image regions and caption words can be applied arbitrarily.

Anderson et al. [53] achieved a good performance by employing a two-stage decoder containing bottom-up attention and top-down attention. Yao et al. [55], Guo et al. [71] and Yang et al. [72] further enhanced the results of the two-stage decoder by introducing scene graphs for detecting the relationship between image regions. Yao et al. [69] achieved better results than [55, 71, 72] by introducing a hierarchical tree and using a tree-LSTM to model the dependency structure.

The best performance in Table I was achieved by Yu et al. [75] and Pan et al. [114]. In [75], Yu et al. used a multimodal transformer that can model three different relations, and the

model was designed in two views aligned and unaligned multi-view visual representation. However, Pan et al. [114] modelled second order interactions through proposing X-linear attention module plugged into transformer. Both of [75] and [114] are Transformer-based attention models which proves that Transformer-based models can achieve better results in comparison with other attention-based mechanisms. The big interest in applying the transformer, as can be seen from Fig. 1, comes from its ability to weight the importance of every input region and its ability to avoid any duplication by employing the attention in a comprehensive way between the input and the output. In addition, it can be parallelized in an effective way.

Employing the attention mechanism in image captioning started from the year 2015 [48] and it is getting more attention from that time since the number of research papers employed the attention is increasing every year as explained in Fig. 1. In addition, the authors have a tendency for using the scene graph with attention models and also great attention is going towards applying the transformer in the image captioning task due to its parallelization nature and better performance. In addition, part of the research in image captioning task recently is going towards applying the attention as multi-layer in order to enhance the predicted words or adaptively determine the number of required attention steps.

TABLE I. COMPARISON BETWEEN THE STATE-OF-THE-ART ATTENTION-BASED CAPTIONING APPROACHES

Ref.	Year	Category of the Attention	Results (C5)			
			B@4	MT	R	C
[48]	2015	Spatial Single Stage	25.0	23.04	-	-
[57]	2016	Semantic Single Stage	31.6	25.0	53.5	94.3
[49]	2017	Spatial Multi-Layer	30.2	24.4	52.4	91.2
[94]	2017	Multi-Layer	32.0	25.8	54.0	102.9
[54]	2017	Adaptive	33.6	26.4	55.0	104.2
[89]	2018	Multi-Layer	26.7	23.4	51.0	84.4
[76]	2018	Transformer	33.3	-	54.8	108.1
[88]	2018	Multi-Layer	33.8	27.0	55.4	106.4
[61]	2018	Two-Stages	34.7	27.1	-	107.2
[93]	2018	Multi-Layer	34.9	26.7	-	108.1
[102]	2018	Guided	35.3	26.7	55.5	107.8
[53]	2018	Two-Stages	36.9	27.6	57.1	117.9
[87]	2018	Multi-Layer	38.1	28.3	58.0	126.1
[55]	2018	Two-Stages with Graph	38.7	28.5	58.5	125.3
[103]	2019	Guided	34.0	26.3	55.2	103.6
[71]	2019	Two-Stages with Graph	37.4	28.2	57.9	123.1
[72]	2019	Two-Stages with Graph	38.5	28.2	58.6	123.8
[97]	2019	Adaptive	38.7	28.6	58.5	128.6
[77]	2019	Transformer	38.9	28.6	58.6	122.1

[69]	2019	Two-Stages with Graph	39.3	28.8	59.0	127.9
[86]	2019	Transformer	39.4	29.1	58.9	126.9
<b>[75]</b>	<b>2019</b>	<b>Transformer</b>	<b>40.4</b>	<b>29.4</b>	<b>59.6</b>	<b>130</b>
[90]	2020	Multi-Layer	28.5	25.3	56.5	92.4
[95]	2020	Adaptive	32.6	27.0	-	-
[66]	2020	Two-Stages with Graph	34.3	27.0	55.5	106.1
[113]	2020	Transformer	38.8	29.0	58.7	126.3
[74]	2020	Transformer	39.7	29.4	59.2	129.3
[50]	2020	Transformer	39.6	29.1	59.2	127.4
<b>[114]</b>	<b>2020</b>	<b>Transformer</b>	<b>40.3</b>	<b>29.6</b>	<b>59.5</b>	<b>131.1</b>
[81]	2021	Transformer	40.0	29.1	59.4	129.4
[82]	2021	Transformer	38.5	28.9	58.6	129.6
[80]	2021	Adaptive	36.3	27.2	56.8	113.3
[98]	2021	Guided	39.8	28.8	59.4	128.3
[99]	2022	Guided	35.9	28.4	57.3	115.9
[85]	2022	Transformer	37.9	28.8	58.1	126.7
[115]	2022	Transformer	39.2	28.8	58.7	125.6
[116]	2022	Transformer	39.9	29.1	59.1	127.8
[117]	2022	Adaptive	38.5	28.3	57.5	120.7

## VI. CONCLUSION AND FUTURE WORK

In this paper, a survey was presented for the attention-based image captioning approaches. Four main categories of the attention-based approaches and their subcategories are summarized. Furthermore, the attention-based approaches were compared on benchmark datasets and popular performance metrics. As discussed in the paper, there is a great improvement in the image captioning task due to using the attention-based models especially using Transformer-based approaches. Although there is an impressive effect of using the attention-based models in image captioning, there is still room for improvement. Faster R-CNN is extensively employed as an encoder because of its ability to get effective detection results. However, training of Faster R-CNN is not a simple task, and it gives unsatisfied results in some cases, like when having images of low resolutions or when the objects are deformed or of small size. So, it may be better if other image encoders are used or when an enhanced version of Faster-RCNN is employed. In addition, another room for improvement can be found in the transformer-based models with introducing new transformer architectures, which may help in improving the quality of the result description.

## REFERENCES

- [1] Xinlei Chen, and C. Lawrence Zitnick. "Mind's eye: A recurrent visual representation for image caption generation," In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 2422-2431.
- [2] Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. "On the properties of neural machine translation:

- Encoder-decoder approaches,” In Association for Computational Linguistics, 2014, pp. 103–111.
- [3] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. “Empirical evaluation of gated recurrent neural networks on sequence modeling,” arXiv preprint arXiv:1412.3555, 2014.
- [4] Bo Dai, Dahua Lin, Raquel Urtasun, and Sanja Fidler. “Towards diverse and natural image descriptions via a conditional GAN,” In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR’17), 2017, pp. 2989–2998.
- [5] Soad Samir, Eid Emary, Khaled El-Sayed, and Hoda Onsi. “Optimization of a pre-trained AlexNet model for detecting and localizing image forgeries,” Information, 2020, 11(5): 275.
- [6] Ahmed Ali Hammam, Mona M. Soliman, Aboul Ella Hassanein. “Real-time multiple spatiotemporal action localization and prediction approach using deep learning,” Neural Networks, 2020, 128: 331–344.
- [7] Elham S. Salama, Reda A.El-Khoribi, Mahmoud E. Shoman, Mohamed A. Wahby Shalaby. “A 3D-convolutional neural network framework with ensemble learning techniques for multi-modal emotion recognition,” Egyptian Informatics Journal, 2021, 22(2): 167–176.
- [8] Navneet Dalal and Bill Triggs. “Histograms of oriented gradients for human detection,” In IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR’05, 2005, 1: 886–893.
- [9] Marie-Catherine De Marneffe, Bill MacCartney, and Christopher D. Manning. “Generating typed dependency parses from phrase structure parses,” In Proceedings of LREC, 2006, 6: 449–454.
- [10] Etienne Denoual and Yves Lepage. “BLEU in characters: Towards automatic MT evaluation in languages without word delimiters,” In Companion Volume to the Proceedings of the 2nd International Joint Conference on Natural Language Processing, 2005, pp. 81–86.
- [11] Jacob Devlin, Hao Cheng, Hao Fang, Saurabh Gupta, Li Deng, Xiaodong He, Geoffrey Zweig, and Margaret Mitchell. “Language models for image captioning: The quirks and what works,” arXiv preprint arXiv:1505.01809, 2015.
- [12] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. “Long-term recurrent convolutional networks for visual recognition and description,” In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 2625–2634.
- [13] Vinyals, Oriol, Alexander Toshev, Samy Bengio, and Dumitru Erhan. “Show and tell: A neural image caption generator,” In Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 3156–3164.
- [14] A. Karpathy and L. Fei-Fei. “Deep visual-semantic alignments for generating image descriptions,” CVPR, 2015.
- [15] Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, Zhiheng Huang, and Alan Yuille. “Deep captioning with multimodal recurrent neural networks (m-RNN),” In International Conference on Learning Representations (ICLR’15), 2015.
- [16] J. Mao, W. Xu, Y. Yang, J. Wang, and A. L. Yuille. “Explain images with multimodal recurrent neural networks,” arXiv preprint arXiv:1410.1090, 2014.
- [17] Hossain, MD Zakir, Ferdous Sohel, Mohd Fairuz Shiratuddin, and Hamid Laga. “A comprehensive survey of deep learning for image captioning,” ACM Computing Surveys (CSUR), 2019, 51(6): 1–36.
- [18] Ryan Kiros, Ruslan Salakhutdinov, and Richard S. Zemel. “Unifying visual-semantic embeddings with multimodal neural language models,” In Workshop on Neural Information Processing Systems (NIPS’14), 2014.
- [19] Andrej Karpathy, Armand Joulin, and Fei Fei F. Li. “Deep fragment embeddings for bidirectional image sentence mapping,” In Advances in Neural Information Processing Systems, 2014, pp. 1889–1897.
- [20] Chen Chen, Shuai Mu, Wanpeng Xiao, Zexiong Ye, Liesi Wu, and Qi Ju. “Improving image captioning with conditional generative adversarial nets,” In Proceedings of the AAAI Conference on Artificial Intelligence, 2019, 33(1): 8142–8150.
- [21] Rakshith Shetty, Marcus Rohrbach, Lisa Anne Hendricks, Mario Fritz, and Bernt Schiele. “Speaking the same language: Matching machine to human captions by adversarial training,” In IEEE International Conference on Computer Vision (ICCV’17), 2017, pp. 4155–4164.
- [22] Zhou Ren, Xiaoyu Wang, Ning Zhang, Xutao Lv, and Li-Jia Li. “Deep reinforcement learning-based image captioning with embedding reward,” In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR’17), 2017, pp. 1151–1159.
- [23] Steven J. Rennie, Etienne Marcheret, Youssef Mroueh, Jarret Ross, and Vaibhava Goel. “Self-critical sequence training for image captioning,” In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR’17), 2017, pp. 1179–1195.
- [24] Jiuxiang Gu, Gang Wang, Jianfei Cai, and Tsuhan Chen. “An empirical study of language CNN for image captioning,” In Proceedings of the International Conference on Computer Vision (ICCV’17), 2017, pp. 1231–1240.
- [25] Siqi Liu, Zhenhai Zhu, Ning Ye, Sergio Guadarrama, and Kevin Murphy. “Improved image captioning via policy gradient optimization of spider,” In Proceedings of the IEEE International Conference on Computer Vision (ICCV’17), 2017, 3: 873–881.
- [26] Shubo Ma and Yahong Han. “Describing images by feeding LSTM with structural words,” In 2016 IEEE International Conference on Multimedia and Expo (ICME’16), IEEE, 2016, pp. 1–6.
- [27] Subhashini Venugopalan, Lisa Anne Hendricks, Marcus Rohrbach, Raymond Mooney, Trevor Darrell, and Kate Saenko. “Captioning images with diverse objects,” In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1170–1178.
- [28] Qi Wu, Chunhua Shen, Peng Wang, Anthony Dick, and Anton van den Hengel. “Image captioning and visual question answering based on attributes and external knowledge,” IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(6), pp. 1367–1381.
- [29] Li Zhang, Flood Sung, Feng Liu, Tao Xiang, Shaogang Gong, Yongxin Yang, and Timothy M. Hospedales. “Actor-critic sequence training for image captioning,” arXiv preprint arXiv:1706.09601, 2017.
- [30] Ryan Kiros, Ruslan Salakhutdinov, and Rich Zemel. “Multimodal neural language models,” In Proceedings of the 31st International Conference on Machine Learning (ICML’14), 2014, pp. 595–603.
- [31] Justin Johnson, Andrej Karpathy, and Li Fei-Fei. “Densecap: Fully convolutional localization networks for dense captioning,” In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 4565–4574.
- [32] Jyoti Aneja, Aditya Deshpande, and Alexander G. Schwing. “Convolutional image captioning,” In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 5561–5570.
- [33] Y. Bin, Y. Yang, J. Zhou, Z. Huang, and H.T. Shen. “Adaptively Attending to Visual Attributes and Linguistic Knowledge for Captioning,” In Proceedings of the 2017 ACM on Multimedia Conference, 2017, pp. 1345–1353.
- [34] S. Qu, Y. Xi, and S. Ding. “Visual Attention Based on Long-Short Term Memory Model for Image Caption Generation,” Control and Decision Conference (CCDC), 2017 29th Chinese, 2017, pp. 4789–4794.
- [35] L. Li, S. Tang, Y. Zhang, L. Deng, and Q. Tian. “GLA: Global-Local Attention for Image Description,” IEEE Trans. on Multimedia, 2018, 20(3): 726–737.
- [36] S. Ye, J. Han, and N. Liu. “Attentive Linear Transformation for Image Captioning,” IEEE Trans. on Image Processing, 2018, 27(11): 5514–5524.
- [37] Cornia, Marcella, Lorenzo Baraldi, Giuseppe Serra, and Rita Cucchiara. “Paying more attention to saliency: Image captioning with saliency and context attention,” ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 2018, 14(2): 1–21.
- [38] Bai, Shuang, and Shan An. “A survey on automatic image caption generation,” Neurocomputing 311, 2018, pp. 291–304.
- [39] Bernardi, Raffaella, Ruket Cakici, Desmond Elliott, Aykut Erdem, Erkut Erdem, Nazli Ikizler-Cinbis, Frank Keller, Adrian Muscat, and Barbara Plank. “Automatic description generation from images: A survey of models, datasets, and evaluation measures,” Journal of Artificial Intelligence Research 55, 2016, pp. 409–442.

- [40] Liu, Xiaoxiao, Qingyang Xu, and Ning Wang. "A survey on deep neural network-based image captioning," *The Visual Computer*, 2019, 35(3): 445-470.
- [41] Staniūtė, Raimonda, and Dmitrij Šešok. "A systematic literature review on image captioning," *Applied Sciences*, 2019, 9(10): 2024.
- [42] Wang, Yiyu, Jungang Xu, Yingfei Sun, and Ben He. "Image Captioning based on Deep Learning Methods: A Survey," *arXiv preprint arXiv:1905.08110*, 2019.
- [43] Khaing, Phyu Phyu. "Attention-Based Deep Learning Model for Image Captioning: A Comparative Study," *International Journal of Image, Graphics and Signal Processing*, 2019, 10(6): 1.
- [44] Zohourianshahzadi, Zanyar, and Jugal K. Kalita. "Neural attention for image captioning: review of outstanding methods," *Artificial Intelligence Review*, 2022, pp. 1-30.
- [45] Bahdanau, Dzmitry, Cho, Kyunghyun, and Bengio, Yoshua. "Neural machine translation by jointly learning to align and translate," *arXiv:1409.0473*, September 2014.
- [46] Ba, Jimmy Lei, Mnih, Volodymyr, and Kavukcuoglu, Koray. "Multiple object recognition with visual attention," *arXiv:1412.7755*, December 2014.
- [47] Mnih, Volodymyr, Hees, Nicolas, Graves, Alex, and Kavukcuoglu, Koray. "Recurrent models of visual attention," In *NIPS*, 2014.
- [48] Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., Bengio, Y. "Show, attend and tell: Neural image caption generation with visual attention," In *International conference on machine learning*, 2015, pp. 2048-2057.
- [49] Chen, L., Zhang, H., Xiao, J., Nie, L., Shao, J., Liu, W., Chua, T.S. "Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning," In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 5659-5667.
- [50] He, Sen, Wentong Liao, Hamed R. Tavakoli, Michael Yang, Bodo Rosenhahn, and Nicolas Pugeault. "Image captioning through image transformer," In *Proceedings of the Asian Conference on Computer Vision*, 2020.
- [51] C. Koch and S. Ullman. "Shifts in selective visual attention: towards the underlying neural circuitry," In *Matters of intelligence*, Springer, 1987, pp. 115-141.
- [52] M.W. Spratling and M. H. Johnson. "A feedback model of visual attention," In *Journal of cognitive neuroscience*, 2004, 16(2): 219-237.
- [53] Anderson, Peter, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. "Bottom-up and top-down attention for image captioning and visual question answering," In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6077-6086.
- [54] Lu, Jiasen, Caiming Xiong, Devi Parikh, and Richard Socher. "Knowing when to look: Adaptive attention via a visual sentinel for image captioning," In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 375-383.
- [55] Yao, Ting, Yingwei Pan, Yehao Li, and Tao Mei. "Exploring visual relationship for image captioning," In *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 684-699.
- [56] Zhou, Dongming, Jing Yang, and Riqiang Bao. "Collaborative strategy network for spatial attention image captioning," *Applied Intelligence* 52, no. 8 (2022): 9017-9032.
- [57] You, Quanzeng, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. "Image captioning with semantic attention," In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4651-4659.
- [58] Gan, Zhe, Chuang Gan, Xiaodong He, Yunchen Pu, Kenneth Tran, Jianfeng Gao, Lawrence Carin, and Li Deng. "Semantic compositional networks for visual captioning," In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 5630-5639.
- [59] Luo, W., Li, Y., Urtasun, R., Zemel, R. "Understanding the effective receptive field in deep convolutional neural networks," In *Advances in neural information processing systems*, 2016, pp. 4898-4906.
- [60] Ren, Shaoqing, Kaiming He, Ross Girshick, and Jian Sun. "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, 2015, 28: 91-99.
- [61] Lu, Jiasen, Jianwei Yang, Dhruv Batra, and Devi Parikh. "Neural baby talk," In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7219-7228.
- [62] Jin, Junqi, Kun Fu, Rungpeng Cui, Fei Sha, and Changshui Zhang. "Aligning where to see and what to tell: image caption with region-based attention and scene factorization," *arXiv preprint arXiv:1506.06272*, 2015.
- [63] Kipf, T.N., Welling, M. "Semi-supervised classification with graph convolutional networks," In: *ICLR*, 2017.
- [64] Y. Li, W. Ouyang, B. Zhou, K. Wang, X. Wang. "Scene graph generation from objects, phrases and region captions," In *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1261-1270.
- [65] Krishna, Ranjay, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen et al. "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *International journal of computer vision*, 2017, 123(1): 32-73.
- [66] Junbo Wang, Wei Wang, Liang Wang, Zhiyong Wang, David Dagan Feng, Tieniu Tan. "Learning Visual Relationship and Context-Aware Attention for Image Captioning," *Pattern Recognition*, 2020.
- [67] Plummer, Bryan A., Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. "Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models," In *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2641-2649.
- [68] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. "Microsoft Coco: Common objects in context," In *European Conference on Computer Vision*, Springer, 2014, pp. 740-755.
- [69] Yao, Ting, Yingwei Pan, Yehao Li, and Tao Mei. "Hierarchy parsing for image captioning," In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 2621-2629.
- [70] Kai Sheng Tai, Richard Socher, and Christopher D Manning. "Improved semantic representations from tree-structured long short-term memory networks," In *ACL*, 2015.
- [71] Guo, Longteng, Jing Liu, Jinhui Tang, Jiangwei Li, Wei Luo, and Hanqing Lu. "Aligning linguistic words and visual semantic units for image captioning," In *Proceedings of the 27th ACM International Conference on Multimedia*, 2019, pp. 765-773.
- [72] Yang, Xu, Kaihua Tang, Hanwang Zhang, and Jianfei Cai. "Auto-encoding scene graphs for image captioning," In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10685-10694.
- [73] Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. "Attention is all you need," In *Advances in neural information processing systems*, 2017, pp. 5998-6008.
- [74] Cornia, Marcella, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. "Meshed-memory transformer for image captioning," In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10578-10587.
- [75] Yu, Jun, Jing Li, Zhou Yu, and Qingming Huang. "Multimodal transformer with multi-view visual representation for image captioning," *IEEE transactions on circuits and systems for video technology*, 2019, 30(12):4467-4480.
- [76] Zhu, Xinxin, Lixiang Li, Jing Liu, Haipeng Peng, and Xinxin Niu. "Captioning transformer with stacked attention modules," *Applied Sciences*, 2018, 8(5): 739.
- [77] Li, Guang, Linchao Zhu, Ping Liu, and Yi Yang. "Entangled transformer for image captioning," In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 8928-8937.
- [78] Herdade, Simao, Armin Kappeler, Kofi Boakye, and Joao Soares. "Image captioning: Transforming objects into words," *arXiv preprint arXiv:1906.05963*, 2019.
- [79] Jiangyun Li, Peng Yao, Longteng Guo, and Weicun Zhang. "Boosted transformer for image captioning," *Applied Sciences*, 2019, 9(16): 3260.
- [80] Chenggang Yan, Yiming Hao, Liang Li, Jian Yin, Anan Liu, Zhendong Mao, Zhenyu Chen, and Xingyu Gao. "Task-adaptive attention for

- image captioning,” IEEE Transactions on Circuits and Systems for Video technology, 2021, 32(1): 43-51.
- [81] Wei Liu, Sihan Chen, Longteng Guo, Xinxin Zhu, and Jing Liu. “CPTR: Full transformer network for image captioning,” arXiv preprint arXiv:2101.10804, 2021.
- [82] Weitao Jiang, Xiyang Li, Haifeng Hu, Qiang Lu, and Bohong Liu. “Multi-gate attention network for image captioning,” IEEE Access, 2021, 9: 69700-69709.
- [83] Kumar, Deepika, Varun Srivastava, Daniela Elena Popescu, and Jude D. Hemanth. “Dual-Modal Transformer with Enhanced Inter-and Intra-Modality Interactions for Image Captioning,” Applied Sciences 12, no. 13 (2022): 6733.
- [84] Sarto, Sara, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. “Retrieval-augmented transformer for image captioning,” In Proceedings of the 19th International Conference on Content-based Multimedia Indexing, pp. 1-7. 2022.
- [85] Dubey, Shikha, Farrukh Olimov, Muhammad Aasim Rafique, Joonmo Kim, and Moongu Jeon. “Label-attention transformer with geometrically coherent objects for image captioning,” Information Sciences, 2022.
- [86] Huang, Lun, Wenmin Wang, Jie Chen, and Xiao-Yong Wei. “Attention on attention for image captioning,” In Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 4634-4643.
- [87] Du, Jiajun, Yu Qin, Hongtao Lu, and Yonghua Zhang. “Attend more times for image captioning,” arXiv preprint arXiv:1812.03283, 2018.
- [88] Zhu, Xinxin, Lixiang Li, Jing Liu, Ziyi Li, Haipeng Peng, and Xinxin Niu. “Image captioning with triple-attention and stack parallel LSTM,” Neurocomputing, 2018, 319: 55-65.
- [89] Wang, Qingzhong, and Antoni B. Chan. “Cnn+ cnn: Convolutional decoders for image captioning,” arXiv preprint arXiv:1805.09019, 2018.
- [90] Yan, Shiyang, Yuan Xie, Fangyu Wu, Jeremy S. Smith, Wenjin Lu, and Bailing Zhang. “Image captioning via hierarchical attention mechanism and policy gradient optimization,” Signal Processing, 2020, 167: 107329.
- [91] Wang, Qingzhong, and Antoni B. Chan. “Gated hierarchical attention for image captioning” In Asian Conference on Computer Vision, Springer, Cham, 2018, pp. 21-37.
- [92] Wang, Weixuan, Zhihong Chen, and Haifeng Hu. “Hierarchical attention network for image captioning,” In Proceedings of the AAAI Conference on Artificial Intelligence, 2019, 33(01): 8957-8964.
- [93] Fang, Fang, Qinyu Li, Hanli Wang, and Pengjie Tang. “Refining attention: a sequential attention model for image captioning,” In 2018 IEEE international conference on multimedia and expo (ICME), 2018, pp. 1-6.
- [94] Liu, Chang, Fuchun Sun, Changhu Wang, Feng Wang, and Alan Yuille. “MAT: A multimodal attentive translator for image captioning,” arXiv preprint arXiv:1702.05658, 2017.
- [95] Deng, Zhenrong, Zhouqin Jiang, Rushi Lan, Wenming Huang, and Xiaonan Luo. “Image captioning using DenseNet network and adaptive attention,” Signal Processing: Image Communication, 2020, 85: 115836.
- [96] Xiao, Fen, Xue Gong, Yiming Zhang, Yanqing Shen, Jun Li, and Xieping Gao. “DAA: Dual LSTMs with adaptive attention for image captioning,” Neurocomputing, 2019, 364: 322-329.
- [97] Huang, Lun, Wenmin Wang, Yaxian Xia, and Jie Chen. “Adaptively aligned image captioning via adaptive attention time,” Advances in Neural Information Processing Systems, 2019, 32: 8942-8 951.
- [98] Ziwei Tang, Yaohua Yi, and Hao Sheng. “Attention-Guided Image Captioning through Word Information,” Sensors, 2021, 21(23): 7982.
- [99] Murad Popattia, Muhammad Rafi, Rizwan Qureshi, and Shah Nawaz. “Guiding Attention using Partial-Order Relationships for Image Captioning,” In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 4671-4680.
- [100] Jia, Xu, Efstratios Gavves, Basura Fernando, and Tinne Tuytelaars. “Guiding the long-short term memory model for image caption generation,” In Proceedings of the IEEE international conference on computer vision, 2015, pp. 2407-2415.
- [101] Yao, Ting, Yingwei Pan, Yehao Li, Zhaofan Qiu, and Tao Mei. “Boosting image captioning with attributes,” In Proceedings of the IEEE international conference on computer vision, 2017, pp. 4894-4902.
- [102] Jiang, Wenhao, Lin Ma, Xinpeng Chen, Hanwang Zhang, and Wei Liu. “Learning to guide decoding for image captioning,” In Thirty-second AAAI conference on artificial intelligence, 2018.
- [103] Sow, Daouda, Zengchang Qin, Mouhamed Niasse, and Tao Wan. “A sequential guiding network with attention for image captioning,” In ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019, pp. 3802-3806.
- [104] Thang Luong, Hieu Pham, and Christopher D. Manning. “Effective approaches to attention-based neural machine translation,” in EMNLP, 2015.
- [105] Mun, Jonghwan, Minsu Cho, and Bohyung Han. “Text-guided attention model for image captioning,” In Proceedings of the AAAI Conference on Artificial Intelligence, 2017, 31(1).
- [106] Zhu, Zhihao, Zhan Xue, and Zejian Yuan. “Topic-guided attention for image captioning,” In 2018 25th IEEE International Conference on Image Processing (ICIP), 2018, pp. 2615-2619.
- [107] Hodosh, Micah, Peter Young, and Julia Hockenmaier. “Framing image description as a ranking task: Data, models and evaluation metrics,” Journal of Artificial Intelligence Research, 2013, 47: 853-899.
- [108] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zh. BLEU: “A method for automatic evaluation of machine translation,” In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics, 2002, pp. 311-318.
- [109] Chin-Yew Lin. “Rouge: A package for automatic evaluation of summaries,” In Text Summarization Branches Out: Proceedings of the ACL-04 Workshop, 2004, vol. 8.
- [110] Satanjeev Banerjee and Alon Lavie. “METEOR: An automatic metric for MT evaluation with improved correlation with human judgments,” In Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, 2005, 29: 65-72.
- [111] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. “Cider: Consensus-based image description evaluation,” In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 4566-4575.
- [112] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. “Spice: Semantic propositional image caption evaluation,” In European Conference on Computer Vision. Springer, 2016, pp. 382-398.
- [113] Guo, Longteng, Jing Liu, Xinxin Zhu, Peng Yao, Shichen Lu, and Hanqing Lu. “Normalized and geometry-aware self-attention network for image captioning,” In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 10327-10336.
- [114] Pan, Yingwei, Ting Yao, Yehao Li, and Tao Mei. “X-linear attention networks for image captioning,” In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 10971-10980.
- [115] Wei, Yiwei, Chunlei Wu, Guohe Li, and Haitao Shi. “Sequential Transformer via an Outside-In Attention for image captioning,” Engineering Applications of Artificial Intelligence 108 (2022): 104574.
- [116] Wang, Chi, Yulin Shen, and Luping Ji. “Geometry Attention Transformer with position-aware LSTMs for image captioning,” Expert Systems with Applications 201 (2022): 117174.
- [117] Wang, Changzhi, and Xiaodong Gu. “Image captioning with adaptive incremental global context attention,” Applied Intelligence 52, no. 6 (2022): 6575-6597.

# Towards an Automatic Speech-to-Text Transcription System: Amazigh Language

Ahmed Ouhnini<sup>1\*</sup>, Brahim Aksasse<sup>2</sup>, Mohammed Ouanan<sup>3</sup>  
Dept. of Computer Science FST, Moulay Ismail University, Errachidia, Morocco<sup>1</sup>  
Dept. of Computer Science FS, Moulay Ismail University, Meknes, Morocco<sup>2,3</sup>

**Abstract**—Various studies inside the domain of research and the development of automatic speech recognition (ASR) technologies for several languages have not yet been published and thoroughly investigated. Nevertheless, the unique acoustic features of the Amazigh language, for example, Amazigh's consonant emphasis, pose many obstacles to the development of automatic speech recognition systems. In this study, we examine Amazigh language voice recognition. We treat the problem by focusing on transitions in vowel and consonant sounds and formant frequencies of phonemes. We present a hybrid strategy for phoneme separation based on energy differences. This includes analysis of consonant and vowel features, and identification methods based on formant analysis.

**Keywords**—Speech recognition system; Amazigh language; analyzing formants and pitch; speech corpus; artificial intelligence

## I. INTRODUCTION

Automatic Speech Recognition (ASR) is used to transcribe human speech captured via a microphone into text that computers can understand in order to enhance human-machine (HM) communication. ASR has long been the subject of intense research. Formant frequencies have been studied for decades. ASR has long been a topic of active investigation. For many years, formant frequencies are believed to be an important factor in recognizing speech phonetic content [1]. To arrive at the stage of recognizing phonemes, we examine a specific case of this problem and focus on the vowels and consonants transition in the Amazigh language and the formant frequencies of phonemes that are important for determining the phonetic content of speech [2]. We trust that this effort, will highlight the importance of consonant-vowel changes and vocal parameter analysis in speech recognition.

We provide a strategy that includes separation of phonemes by differences in energy between consonants and vowels, vocal characteristics processing of phonetics units, and a recognizer algorithm based on formants. Formant analysis methods focus on associating physical aspects with phonology. This method determines speech types by analyzing linguistically distinct features of the speech.

The rest of this article is organized as follows: Section II describes the characteristics of the human voice and the Amazigh language. The remainder of Section III describes some mathematical and engineering techniques for language modeling, followed by a discussion of proposed phoneme recognition methods. Section IV provides further insight into the results and Section V concludes the article.

## II. CHARACTERISTICS OF SPEECH AND PHONETICS

### A. Human Ear and Acoustic Sound

Sound constitutes a wave that propagates in a material environment like small variations of pressure. This is perceptible via human ears at frequencies ranging from 20 Hz to 20 kHz. Nonetheless, the phonetic information is judged to be less than 10 kHz [3].

Due to the way sound signals are sifted, we modify the sufficient range because ears are not sensitive to stage distortion. This allows us to focus exclusively on complementary application modules.

### B. Human Voice

The human voice is a collaboration of breathing and multiple phonatory organs. In voiced phonemes, sound is first produced by the vibration of the vocal cords [4]. It is manipulated differently depending on the cavities it passes through, primarily the pharynx and mouth. These cavities act as resonators, increasing frequencies corresponding to the resonant frequencies of specific phonemes. These enhanced frequencies are known as "formants" and are the features that phonologists search in spectrograms to identify phonemes being pronounced [5].

The Fig. 1 shows the formants (F1, F2, and F3) superimposed on spectrogram of speech signal « he took holidays » showing the alternating voiced and unvoiced sounds. In the voiced situation, a formant structure is presented.

### C. Phonology and Phonetic

Phonetic by definition is study of phonetic units, the smallest particular phonetic unit being regularly defined as a phoneme. The opposition between the terms bath and bread, well, suggests that [b] and [p] are phonemes. In general, we can classify them as follows: classes and subclasses, more significant of which is "vowel" and "consonant".

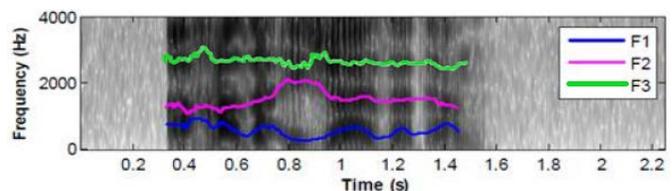


Fig. 1. Illustration of formants F1, F2, F3 of speech signal "he took holidays".

\*Corresponding Author.

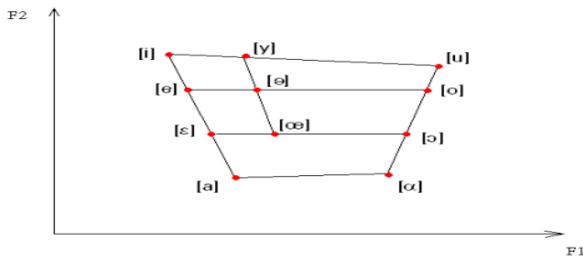


Fig. 2. Categorization of vowels using the vocal trapezoid.

Relatively vowels are lengthy-duration sounds with a flows and with considerable frequency characteristics consistency during times: in absence of highly pronounced prosodic features, formants appear horizontal on the spectrogram [6]. The vowel trapezoid, shown in Fig. 2 illustrates their location in the planes specified by initial both formants, indicating articulation position of language.

However, consonants represent phonemes that encounter an obstacle when articulated (such as labial vowels, toothy teeth, palate closure in [k], etc.). In comparison, they are considerably shorter than vowels and significantly variable in length during time. Could be sonorous or loud. In the resonant scenario, only current formants are present, see the Fig. 3.

D. Speech Signal Frequency Parameters

The bandwidth of voice signal is much larger than the telephone bandwidth (4 kHz) and includes all information's necessary to know to decode human voice.

The fundamental frequency refers to the speed of opening and closing of the vocal chords during phonation. Its value is proportional to the individual's phonatory system size [7]. Voice frequency vary between 80 and 600 Hz based on age and gender.

The spectrogram is a representation in three dimensions, where the X-axis represents time, the Y-axis represents frequency, and the Z-axis represents frequency levels (symbolized by gray levels). Fast Fourier transform (FFT) with sliding window is applied to acquire the voice signal.

E. Amazigh Language

The Amazigh language, commonly called Tamazight or Berber, is one of humanity's earliest languages. Now it extends from the Red Sea to the Canary Islands and from Niger in the Sahara to the Mediterranean Sea, including the northern section of Africa.

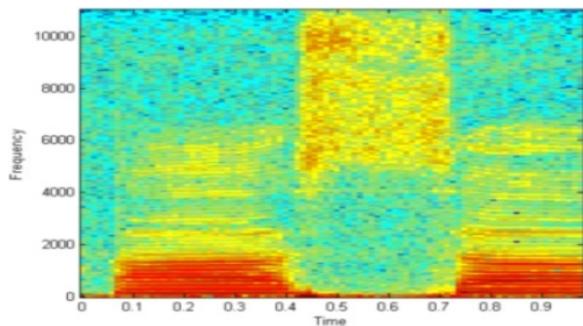


Fig. 3. Spectrogram of the syllable (asa).

In Morocco, Amazigh language is classified into three main regional's varieties, based on historical, geographical, and sociolinguistic factors: Tarifite in the north, Tamazight in Central Morocco and the south-east, and Tachelhite in the south-west and the High Atlas. Even though half of the Moroccan population are Amazigh speakers, the Amazigh language has been reserved exclusively for informal and familial domains: Boukous (1995) [8]. In the last decade, with royal generosity, the language was institutionalized and included in Moroccan school system.

Since February 2003, Morocco's official Amazigh alphabet system, known as Tifinaghe-IRCAM, has been used in Moroccan school programs and Amazigh historical studies. The system uses the alphabet describe in Table I [9]:

TABLE I. AMAZIGH LANGUAGE ALPHABET

<b>27 Consonants</b>	<i>labial</i>	ⵍ, ⵍ, ⵏ
	<i>dental</i>	ⵜ, ⵏ, ⵍ, ⵍ, ⵏ, ⵏ, ⵏ, ⵏ
	<i>alveolar</i>	ⵏ, ⵏ, ⵏ, ⵏ
	<i>palatal</i>	ⵍ, ⵏ
	<i>velar</i>	ⵏ, ⵏ
	<i>labiovelar</i>	ⵏ, ⵏ
	<i>uvular</i>	ⵏ, ⵏ, ⵏ
	<i>pharyngeal</i>	ⵏ, ⵏ
	<i>laryngeal</i>	ⵏ
<b>2 Semi-consonant</b>		ⵏ and ⵏ
<b>4 Vowels</b>	<i>full vowel</i>	ⵏ, ⵏ, ⵏ
	<i>Neuter vowel (schwa)</i>	ⵏ it has a specific status in phonology Amazigh.

The correct writing of words in the Latin letters closely resembles phonetic transcription and correctly conveys their pronunciation, includes twinned and vowel sounds.

Examples: /illa/ → « illa » → « il existe », « il est ». [10]

The pronunciation of Amazigh language varies from region to region [11].

III. TECHNOLOGIES AND METHODS

A. Fourier Analysis

Transform of Fourier permits a time-frequency processing at a resolution suitable for speech signals that are quasi-stationary on intervals of 10-100 ms.

1) *Fourier transform*: We are dealing with the pre-Hilbert sets of square integrables function  $L^2(\mathbb{R})$ , and the orthogonal family of sine functions  $e_f: t \rightarrow e^{j2\pi t} / f \in R$  (we restrict ourselves physically in concret pulses)). Preparing the projection portion of each sinusoid provided by the scalar product, and not ignoring the complex conjugate with respect to the second number, the Fourier transform  $H(f)$  (or  $FT(g(t))$ ) of a functions  $g(.)$  allows projection over the vector space through which the sinusoids:

$$FT(f) = \langle g(t) | e_f \rangle = \int_{-\infty}^{+\infty} g(t) e^{-j2\pi ft} dt \quad (1)$$

Where the variables  $t$  and  $f$  refer to time and frequency, respectively.  $FT(f)(t)$  is the transform of  $g(.)$ .

These transform would be employed in our study to investigate the contributions of each frequency range in the speech signal more qualitatively by evaluating the spectrogram, (Ohm's rule states that human ear is insensitive to the acoustic signal's phase) [12].

2) *Discrete fourier transformation*: The signal has now been expressed by sampling taken evenly throughout time by sampling continuous time  $\{x(n)/n \in [0, N - 1]$

To avoid edge effects, we convert them into an  $N$ -period signal. Following that; the discrete transform expressed as:

$$F(k) = \sum_{n=0}^{N-1} x(n)e^{-\frac{2i\pi kn}{N}} \quad (2)$$

The frequency observations is related to the  $\frac{k}{N}$  factor for  $k \in [0, \frac{N}{2}]$  If  $f_e$  represents the sample frequency, so  $= f_e \times \frac{k}{N}$ . It should be emphasized that because the received information is restricted to half of the period  $[0, N - 1]$ , this transformation is redundant. Indeed:

$$\begin{aligned} \forall k \in Z \quad F(N - k) &= \sum_{n=0}^{N-1} x(n)e^{\frac{2i\pi kn}{N}} e^{-2i\pi n} \\ &= \sum_{n=0}^{N-1} x(n)e^{\frac{2i\pi kn}{N}} = \overline{F(k)} \end{aligned} \quad (3)$$

As a result, modules are similar, with the term simply having the reverse indication. If we wish to evaluate the signal's about frequency scale of 10 kHz, we need to use a sampling frequency with 20 kHz [13]. Within reality, we employ the FFT (Fast Fourier Transform), which has a computational cost of  $O(N \log_2(N))$  rather than  $O(N^2)$  for straight computing, also use this redundancy to improve the computation.

### B. Convolution and Transformation

The convolution product in continuous-time of two functions is given as:

$$\forall t \in R, (f * h)(t) = \int_{-\infty}^{+\infty} f(\tau)h(t - \tau)d\tau \quad (4)$$

Convolution operators are commutative. Also, since there are associative equations, the following two are valid:

$$\forall t \in R, (f * h)(t) = (h * f) \quad (5)$$

As a consequence of commutativity:

$$((h * g) * f)(t) = ((h * f) * g)(t) \quad (6)$$

The Fourier transform of the two functions ordinary product is the convolution product of Fourier transforms. In addition, the Fourier transform of the two functions convolution product is the usual product of Fourier transforms:

$$\begin{cases} FT(g * h) = FT(g) \times FT(h) \\ FT(g \times h) = FT(g) * FT(h) \end{cases} \quad (7)$$

This finding is also valid for discrete cyclic representations. It will be applied in formants analysis as a result of the speech signal modeling adopted.

### C. Windowing Issue

Practically, in addition to the  $x$ -signal discrete, the observation time of  $2\tau$  has over. Consequently, we see the signal convolution using a window function [14]:

$$\Pi_{\tau}(t) = \begin{cases} 1 & \text{if } t \in [-\tau, \tau] \\ 0 & \text{if not} \end{cases} \quad (8)$$

As according (7), Fourier transformation of  $s(t) \times \Pi(t)$  being a convolution product between window and signal, following equation give gate function:

$$\int_{-\infty}^{+\infty} \Pi(t)e^{-i\omega t} dt = \int_{-\tau}^{\tau} e^{-i\omega t} dt = 2\tau \text{sinc}(\omega\tau) \quad (9)$$

This cardinal sinus has a central lobe with a width of  $2/\tau$ . As a consequence, when the observation period approaches zero, the spectrum expands. To resolve this issue, we use a zero-energy concentration window that restricts this phenomena. A Hamming window was used in our study [15], [16].

$$H(n) = 0.54 - 0.46 \times \cos(2\pi \frac{n}{N-1}) \quad (10)$$

### D. Formants and Pitch Examination

1) *Decoding problem from acoustic to phonetic*: Due to the continuous nature of speech signals, it is difficult to identify different linguistic units such as words, syllables and phonemes in the recorded signal. This problem is known as phonetic acoustic decoding. [17], [28]. We used a process allowing us to identify the transitions between consonants and vowels. The syllables corresponding to this case are available to qualify [18].

2) *Vowel-consonant (VC) and Consonant-vowel (CV) Transitions Detection*: Digital filtering is employed first to remove as much background noise as feasible [19].

The flowchart in Fig. 4 illustrates the identification of vowel-consonant and consonant-vowel transitions.

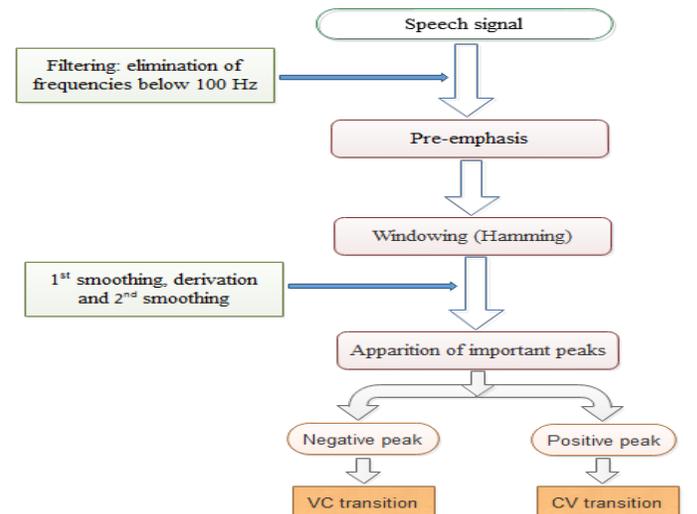


Fig. 4. CV and VC flowchart for detecting transitions [20].

3) *Formants*: Minimums and maximums presents in the vocal signal spectrum correlate to resonance and anti-resonance tract vocal, also named formants and anti-formants.

In general, the frequencies of analysed formants are: F1 is 200-900 Hz, F2 is 500-2500 Hz, F3 is 1500-3500 Hz, and F4 is 2500-4600 Hz.

The formants (F1, F2, F3, and F4) used in this study, for vowels and consonants found in word that recovered in order to differentiate the phoneme formants [21].

Formant analysis is used to identify consonants and vowels. Before formants can be recognized, they must be processed to make them clearer [22].

The chart in Fig. 5 represents this preprocessing.

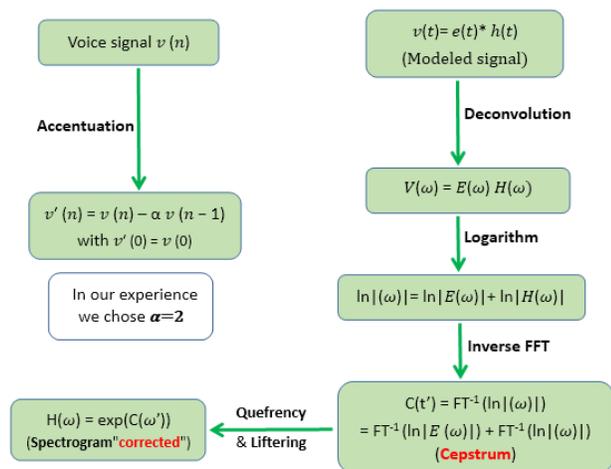


Fig. 5. Diagram for parameter preprocessing.

Once spectrum corrected was acquired, we search for every formant's in a carefully selected frequency band to increase the probability of finding the formant with the highest average amplitude.

Tables II and III show the predicted formant values (in Hertz) for a male voice. The extreme values of each formant have been specified.

TABLE II. A MAN'S VOICE'S ESTIMATED VOWEL FORMANT VALUE

Voyelle	Latin correspondence	F1	F2	F3	F4
[ɛ]	[i]	250	2250	2980	3280
[ɔ]	[u]	420	2050	2630	3340
[ə]	[a]	760	1450	2590	3280

TABLE III. A MAN'S VOICE'S ESTIMATED CONSONANT FORMANT VALUE

Consonne	Latin correspondence	F1	F2	F3	F4
[C]	[m]	300	1300	2300	2770
[l]	[n]	350	1050	2300	3470
[ʃ]	[l]	360	1700	2500	3300
[O]	[r]	550	1300	2300	2700

After the formant values have been obtained, we calculate the distance with every formant values in memory. Formant distance is strictly Euclidean; between both phonemes A and B, it is calculated as follows:

$$\text{dist}(A, B) = \sqrt{\sum_{i=1}^4 (F_A^i - F_B^i)^2} \quad (10)$$

4) *Pitch*: Pitch is a crucial component of human voice and widely recognized as perceptual fundamental of sound that is strongly attached to frequency and can be related to the vocal cord's vibration fundamental frequency, permitting audio frequency recognition. It is among the most essential auditory features of sounds, as well as quality and loudness [23], [30].

We used the "Get Pitch" command to extract the pitch, with set the pitch floor to 75 Hz, and set the pitch ceiling to 500 Hz.

E. Measurement Tools and Corpus

1) *Tools*: Phoneticians and academics utilize the open source program PRAAT [24] to identify various phonetic properties of speech. It is a very efficient software for analysing and recreating acoustic speech signal [25].

For collecting all of the characteristics presented in this study, wav files were registered and analyzed by PRAAT (see Fig. 6).

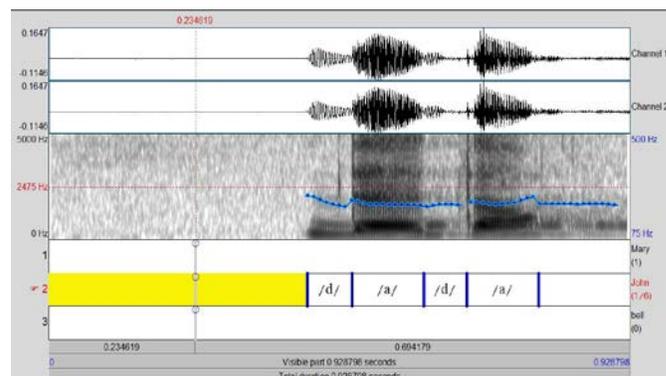


Fig. 6. Manual segmentation of the /dada/ syllable for the CVCV context.

2) *Measurements*: The foundations for measuring voice signals acoustically are pitch and the four formants, which are widely utilized as indicators of perceived speech quality [26]. The Table IV shows the Amazigh vowels used in this work.

TABLE IV. THE VOWELS AMAZIGH USED IN THIS STUDY

Tifinagh	English transcription	Latin correspondence	Arabic transcription	IPA
ⵢ	YA	A	يا	æ
ⵣ	YI	I	يى	i
ⵣⵓ	YO	U	يو	u

3) *Preparation of the corpus*: Ten persons (five women and five men) are chosen among a vocal database of Amazigh

people in Morocco comes from various regions with no distinctive geographical distribution. Age was used to coordinate subjects. The average age of the women was 35, ranging from 23 to 50 years. The men in this group range in age from 25 to 50 years, with an average of 36 years. Each speaker repeated the process 10 times. The total amount of evaluated words (10 speakers x 8 words x 10 repetitions), giving us 800 files to examine.

Our objective is to analyse the consonants, semi consonants, and vowels that are pronounced by studying the important voice parameters [27]. We manually recovered the vowels A, I, and U from Krad, Tanmirt, and Ayur words spectrograms. Based on the spectrogram of words Aghrum, Attas, and Tazalit, R, T, and Z are obtained consonants. More information about the database is shown in Table V.

TABLE V. RECORDING PARAMETERS USED IN THE CORPUS AMAZIGH PHONEMES PREPARATION

Parameter	Value
Sampling rate	22.05 kHz
Quantization	16 bits
Duration	2 second / syllabe
Wave format	Mono, wav
Corpus	10 Amazigh words
Speaker	10 (5 females + 5 males)
Accent	Moroccan Tamazight

4) *Materials*: In this study, we use a microphones and a computer having 8 GB of RAM and an Intel Core i7 processor running at 2.5 GHz. Our experience indicates that Windows 10 LTSB is the prevalent operating system. In a silent room, the microphone were placed between 4 and 10 centimeters from the individual's lips. We recorded the wav file with the parameters shown in Table V.

#### IV. RESULTS AND DISCUSSION

Fig. 7 represents our approach to determining the acoustical power of syllable [ara]. The Fig. 8 give the temporal derivatives of acoustical power shown in Fig. 7.

##### A. Authors and Affiliations

After a series of studies [29], it has been experimentally estimated that the transition occurs near the peak where the signal has lost or gained 66 percent in extreme difference of intensity while comparing both phonemes. Crosses appear on the chart to indicate the transitions.

The algorithm for phoneme separation is very effective as follows:

- The initial and final moments of silence were deleted;
- The speakers didn't blow into the microphone during recording, causing audio signal saturation and resulting in extremely big peak which the program interprets like a transition;

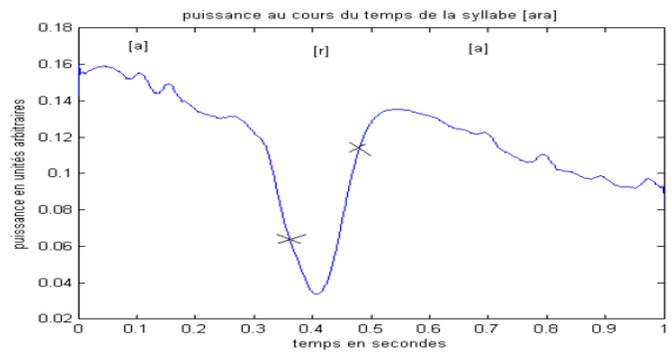


Fig. 7. Acoustical power of the word [ara].

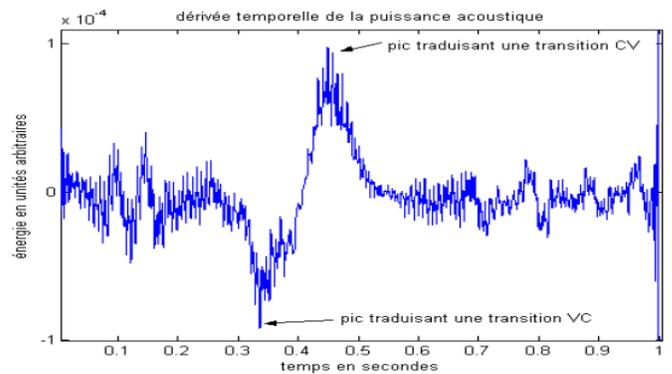


Fig. 8. Derivative of word's acoustical energy [ara].

This method achieves a 75% success rate, which is dependent on parameters such as geographic place, recorded material, as well as verbal difficulties of speakers.

#### V. CONCLUSION

This study examined the difficulty of automatic speech recognition for the Amazigh language. In specifically, we attempted to extract the vowels and consonants from the Amazigh voice signal. It is now possible to develop an Amazigh corpus to complement those that already exist. This approach opens the door to the socioeconomic growth of the Amazigh community in Morocco.

This methodology to voice recognition enabled us to detect and to exploit by solutions implemented, a number of phonetic and spectral characteristics of the Amazigh voice signal. Better identification of the issue parameters (accentuation coefficients, quefrencce cut cepstrum, etc.) in addition to a more precise analysis of formant transitions and trajectories and a main aspects of prosody in the speech are mostly feasible strategies to produce improved outcomes. In addition, speakers of a language like Amazigh should be proficient in both the language and the use of Information Technology resources.

#### REFERENCES

- [1] B. H. Juang et L. R. Rabiner, « Automatic Speech Recognition – A Brief History of the Technology Development », Ga. Inst. Technol. Atlanta Rutgers Univ. Univ. Calif. St. Barbara, p. 24, 2004.
- [2] D. Gerhard, « Pitch Extraction and Fundamental Frequency: History and Current Techniques », Tech. Rep. Regina Dep. Comput. Sci. Univ. Regina, p. 23, 2003.

- [3] T. W. Parsons, « Separation of speech from interfering speech by means of harmonic selection », *J. Acoust. Soc. Am.*, vol. 60, no 4, p. 911-918, oct. 1976, doi: 10.1121/1.381172.
- [4] K. Samudravijaya, « Modeling Natural Language for Automatic Speech Recognition », *Tata Institute Fundam. Res. Homi Bhabha Road Mumbai India*, p. 8.
- [5] J. Li, L. Deng, R. Haeb-Umbach, et Y. Gong, « Fundamentals of speech recognition », in *Robust Automatic Speech Recognition*, Elsevier, 2016, p. 9-40. doi: 10.1016/B978-0-12-802398-3.00002-7.
- [6] T. Koizumi, M. Mori, S. Taniguchi, et M. Maruya, « Recurrent neural networks for phoneme recognition », in *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP '96*, Philadelphia, PA, USA, 1996, vol. 1, p. 326-329. doi: 10.1109/ICSLP.1996.607119.
- [7] N. Dave, « Feature Extraction Methods LPC, PLP and MFCC In Speech Recognition », *Int. J. Adv. Res. Eng. Technol.*, vol. 1, no 6, p. 5, 2013.
- [8] A. Boukous, « Phonologie amazighe », *Inst. R. Cult. Amaz. Rabat Maroc*, p. 445, 2009.
- [9] F. Ataa Allah et S. Boulaknadel, « Amazigh Search Engine: Tifinaghe Character-Based Approach », *Int'l Conf Inf. Knowl. Eng.*, p. 5, 2010.
- [10] M. Makhoul, D. Legros, et B. Marin, « Influence de la langue maternelle kabyle et arabe sur l'apprentissage de l'orthographe française », *Univ. Mouloud Mammeri Tizi Ouzou IUFM Crêteil*, p. 7, 2006.
- [11] H. Satori et F. El Haoussi, « Investigation Amazigh speech recognition using CMU tools », *Int. J. Speech Technol.*, vol. 17, no 3, p. 235-243, sept. 2014, doi: 10.1007/s10772-014-9223-y.
- [12] R. Dufour, « Transcription Automatique de la Parole Spontanée », *Inform. Cs Univ. Maine Fr. Tel-00595465*, p. 190, 2010.
- [13] S. K. Saksamudre, P. P. Shrishrimal, et R. R. Deshmukh, « A Review on Different Approaches for Speech Recognition System », *Int. J. Comput. Appl.*, vol. 115, no 22, p. 23-28, avr. 2015, doi: 10.5120/20284-2839.
- [14] H. Hosni, Z. Sakka, A. Kachouri, et M. Samet, « Étude de la Paramétrisation RASTA PLP en vue de la Reconnaissance Automatique de la Parole Arabe », *5th IEEE Int. Conf. Sci. Electron. Technol. Inf. Telecommun. Tunis.*, p. 7, 2009.
- [15] M. Agrawal et T. Raikwar, « Speech Recognition Using Signal Processing Techniques », *Int. J. Eng. Innov. Technol.*, vol. 5, no 8, p. 4, 2016, doi: 10.17605/osf.io/zab7g.
- [16] T. W. Parsons, *Voice and speech processing / Thomas W. Parsons*. New York: McGraw-Hill, 1987.
- [17] A. Abenaou, F. Ataa Allah, et B. Nsiri, « Vers un système de reconnaissance automatique de la parole en amazighe basé sur les transformations orthogonales paramétrables », *Asinag*, no 9, p. 133-145, 2014.
- [18] B. Lecouteux, « Reconnaissance Automatique de la Parole guidée par des transcriptions a priori », *Inform. Lang. CsCL Univ. D'Avignon Pays Vaucluse Fr. Tel-01381704*, p. 170, 2008.
- [19] S. El Ouahabi, M. Atounti, et M. Bellouki, « Toward an automatic speech recognition system for amazigh-tarifit language », *Int. J. Speech Technol.*, vol. 22, no 2, p. 421-432, juin 2019, doi: 10.1007/s10772-019-09617-6.
- [20] A. Ouhini, B. Aksasse, et M. Ouanan, « Phonemes Recognition Using Formant Analysis in the Case of Consonant Vowel Transition Case "Amazigh Language" », in : *Int'l Conf. Advanced Intelligent Systems for Sustainable Development (AI2SD'2020)*, vol. 1417, Springer International Publishing, 2022, p. 348-358. doi: 10.1007/978-3-030-90633-7\_30.
- [21] C. Gendrot, « Analyses formantiques automatiques de voyelles orales : évidence de la réduction vocalique en langues française et allemande », *MIDL Paris*, p. 6, 2004.
- [22] W. Xiong, L. Wu, F. Alleva, J. Droppo, X. Huang, et A. Stolcke, « The Microsoft 2017 Conversational Speech Recognition System », in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, avr. 2018, p. 5934-5938. doi: 10.1109/ICASSP.2018.8461870.
- [23] E. V. Bonzi, G. B. Grad, A. M. Maggi, et M. R. Muñoz, « Study of the characteristic parameters of the normal voices of Argentinian speakers », *Pap. Phys.*, vol. 6, p. 060002, juill. 2014, doi: 10.4279/PIP.060002.
- [24] M. Labied et A. Belangour, « Automatic Speech Recognition Features Extraction Techniques: A Multi-criteria Comparison », *Int. J. Adv. Comput. Sci. Appl.*, vol. 12, no 8, 2021, doi: 10.14569/IJACSA.2021.0120821.
- [25] J. Kreiman et B. R. Gerratt, « Perception of aperiodicity in pathological voice », *J. Acoust. Soc. Am.*, vol. 117, no 4, p. 2201-2211, avr. 2005, doi: 10.1121/1.1858351.
- [26] R. Rehman, K. Bordoloi, K. Dutta, N. Borah, et P. Mahanta, « Feature Selection and Classification of Speech Dataset for Gender Identification: A Machine Learning Approach », *Journal of Theoretical and Applied Information Technology*, Vol. 98, no 22, p. 11, nov. 2020.
- [27] L. Besacier, V.-B. Le, E. Castelli, S. Setherey, et L. Protin, « Reconnaissance Automatique de la Parole pour des Langues peu Dotées : Application au Vietnamien et au Khmer », *TALN Dourdan*, p. 12, 2005.
- [28] Z. Yin, « Training & Evaluation System of Intelligent Oral Phonics Based on Speech Recognition Technology », *Int. J. Emerg. Technol. Learn. IJET*, vol. 13, no 04, p. 45, mars 2018, doi: 10.3991/ijet.v13i04.8469.
- [29] F. D. Wang, X. Wang, et S. Lv, « An Overview of End-to-End Automatic Speech Recognition », *Symmetry*, vol. 11, no 8, p. 1018, août 2019, doi: 10.3390/sym11081018.
- [30] F. Jiao, J. Song, X. Zhao, P. Zhao, et R. Wang, « A Spoken English Teaching System Based on Speech Recognition and Machine Learning », *Int. J. Emerg. Technol. Learn. IJET*, vol. 16, no 14, p. 68, juill. 2021, doi: 10.3991/ijet.v16i14.24049.

# Graphical User Interfaces Generation from BPMN (Business Process Model and Notation) via IFML (Interaction Flow Modeling Language) up to PSM (Platform Specific Model) Level

Abir Sajji<sup>1</sup>, Yassine Rhazali<sup>2</sup>, Youssef Hadi<sup>3</sup>

Computer Research Laboratory-Faculty of Science, Ibn Tofail University, Kenitra, Morocco<sup>1,3</sup>  
Information and Communication Systems Engineering Research Group, Higher School of Technology, Meknes, Morocco<sup>2</sup>

**Abstract**—The fundamental concept behind the MDA (Model Driven Architecture) approach is the development of many models, first the Computation Independent Model (CIM), then the Platform Independent Model (PIM), and lastly the Platform Specific Model (PSM) for the concrete implementation of the system. Web applications are just one example of customized software that is now being developed at an increasing rate. Interaction Flow Modeling Language (IFML) was developed to represent the front end of any program that necessitates a powerful interaction with a user through the use of an interface, regardless of the technical details of its implementation. There are various modeling tools for IFML; the Webratio tool is an illustration that facilitates the generation of the entire web application. This article discusses the model transformations in the MDA's approach, starting from the CIM level up to the PSM level through the PIM level. To begin, we created the Business Process Model and Notation (BPMN) and IFML metamodels in Eclipse tool, we created also the BPMN model, and we get the IFML model by applying the shift rules in Atlas Transformation Language (ATL). Finally, we generated the application using a standard tool that implements IFML Webratio tool. A CRUD (Create, Read, Update, and Delete) features for the after-sales service case study were provided to illustrate the conversion strategy from the CIM level via the PIM level to the PSM level.

**Keywords**—MDA (Model Driven Architecture); CIM (Computation Independent Model); PIM (Platform Independent Model); PSM (Platform Specific Model); Model transformations; Graphical User Interfaces; BPMN (Business Process Model and Notation); IFML (Interaction Flow Modeling Language); Webratio tool

## I. INTRODUCTION

The core idea behind the MDA's methodology is the use of models at various points in an application's development lifecycle. MDA promotes requirements development (at the CIM level), design and analysis (at the PIM level), and eventually code (PSM level). MDA's primary goal is to create models that are independent of technological aspects to generate all application code automatically and significantly boost productivity [1].

It has been increasingly crucial to deliver tools and processes for web application development. Because of how complicated are the interfaces, the increased demand for Web

applications, and businesses' overall need to create applications quickly in recent years.

Model-driven architecture's rise demonstrates its value to engineering and its effectiveness in raising application quality, while reducing development time and boosting productivity. [2].

Recently, a new web engineering technology named Interaction Flow Modeling Language (IFML) was suggested for use in developing web and mobile applications [3]. It is an Object Management Group (OMG) standard that makes it easier to describe graphical User Interfaces for desktop computers, laptops, mobile phones, and tablets without regard to the platform on which they are used.

Several tools have been suggested for creating IFML models and codes, including the Webratio tool [4]; a model-driven project with code generators that can create functional applications from IFML models.

Our main goal of this article is to automatically generate graphical user interfaces from PIM to PSM using the Webratio tool. In this article, we are going to treat the model transformations in the MDA's approach, starting early from the CIM to PIM up to the PSM level.

We created in the first step the BPMN and IFML metamodels at Eclipse tool, the business process model of a case study named the after-sales service using the BPMN notation created also, then applied the transformation rules in ATL language [5]. We obtain the IFML model at the PIM level into Eclipse tool, finally, we import the IFML model in .xmi format [32] into the Webratio tool to get graphical user interfaces (GUI) code.

The key benefit of our work is that it uses two standards BPMN and IFML as well as a powerful commercial tool Webratio. To cover every step of the model transformations process in the MDA's approach, starting from the CIM level through the PIM level and up to the PSM level.

The article is structured as follows: in Section II we go into the fundamental techniques that underpin our strategy in the theoretical background, Section III presents the related works, in Section IV we will present our approach by describing the

CIM level to the PIM level transformation rules also the representation of the two metamodels BPMN and IFML. Section V provides a case study of the after-sales service illustrating our methodology starting from the CIM level to the PIM level up to the PSM level, a discussion and limitations are presented in Section VI, and we finish by outlining the ongoing and future work in Section VII.

## II. THEORETICAL BACKGROUND

### A. BPMN

The Object Management Group (OMG) established the Business Process Model and Notation (BPMN), a standard that gives businesses the ability to graphically describe and analyze their internal business processes and consistently communicate these processes. Additionally, the graphical notation makes it easier to understand how business transactions and collaborations across companies are performed [13]. A BPMN model does not accurately depict the system's functional behavior. BPMN's fundamental conceptual primitives are tasks, events, gateways, pools, lanes, and flows [14].

### B. IFML

The OMG developed the Interaction Flow Modeling Language (IFML), to define interaction flow models. These models are employed to explain user-application interaction on the front end.

IFML was created by developers with more than ten years of experience with Webratio tool and WebML. The OMG adopted it as a standard in March 2013 [15].

IFML offers multiple advantages to the design of UIs for desktop, mobile, and web applications. One of the five artifacts the IFML specification provides is IFML visual syntax [4]. The latter, however, is simplified and well-known to developers, which discusses the development of various tools for creating IFML diagrams.

### C. Webratio Tool

Some technologies are designed for IFML model creation and code production for the creation of web and mobile applications. Webratio tool [16] focuses on rapid developments using model-driven development and code generation.

It is a commercial web development tool. When the Webratio tool was launched in 2001, it was built using WebML [17]. Later, the Webratio tool used IFML to replace WebML.

The entire process of modeling a web application is supported by the Webratio tool. Typically, one begins by developing the application's data model, or "domain model". This model defines all the data types that the web application will use.

Site views can be made once the domain model has been established. Each of these site views includes an IFML model. The developer can generate the designed web application after modeling it in IFML.

Last but not least, Webratio tool enables the direct deployment of the web application to a remote server.

As a tool, Webratio facilitates IFML web development from conceptual modeling through execution and also enables the entire web system generation with a click when the created IFML models are valid.

## III. RELATED WORK

We relate the previous works that discuss the application of the model-driven development approach to obtain graphical user interfaces.

The authors in [5] represent a methodology that permits a semi-automatic conversion from the CIM level to the PIM level using the MDA's approach, BPMN, and IFML standards. Several rules for semi-automating the conversion from the CIM level to the PIM level were developed. To achieve this an order management case study was provided to demonstrate the transformation strategy.

In [6] a technique for the model-driven creation of a Graphical User Interface for Internet Applications utilizing IFML was described. Model-driven engineering-related frameworks and technologies were used by the authors.

The construction of a prototype Qt/Taurus code generator based on the IFML standard and appropriate modeling tools, which are expanded to provide platform-specific code generation was described by the researchers in [7].

The goal is to enable low-code development in SKA GUI design, increasing the effectiveness, reliability, and coherency of the UI that is created.

A basic GUI use case is provided to fully illustrate the software development life cycle, starting with requirements and incorporating IFML modeling, Qt/Taurus automatic coding, interface evaluation, and validation.

The authors of this article [8] explored the modeling tools for IFML and presented a comparative analysis while taking into account several criteria. In this study, IFMLEdit.org, Webratio tool, IFML Editor, and MIA-studio were all studied and compared as potential IFML tools. Each tool has advantages and disadvantages.

This paper [9] suggested a modeling approach for creating user interfaces that are based on IFML. First, it describes the benefits of the model shifting process, the Webratio tool, and the extension of IFML to web applications.

Second, a detailed method for mapping IFML to the design environment is provided. The authors presented IFML for the web application, which can fully satisfy the expectations of the user. Its primary benefit is that it makes creating the project's front-end interface simple and quick.

AutoCRUD is a Webratio tool plug-in that aims the development of CRUD operations automatically by producing IFML specifications, the plug-in raises the efficiency of Web developers according to Rodriguez-Echeverria et al. [10].

In this work [11] authors provided a new method for creating the user interface for mobile applications, and implementing it with the Android operating system.

By establishing a language for the creation of graphical user interfaces, the Technology Neutral DSL (Domain-specific language) is designed to be cross-compiled to generate native code for a diversity of platforms.

The development of software applications incorporating a multi-experience User Interface is explored in this study through model-driven approaches [12].

The authors explained how elevating the abstraction level at which these interfaces are defined allows for quicker development, better deployment, and better integration of each interface with the rest of the software system and any other interfaces it may need to cooperate with.

They provide a new Domain Specific Language (DSL) for describing various CUI kinds and demonstrate how this DSL may be a component of an integrated modeling environment that can explain the relationships between the modeled CUIs and the other models of the system.

#### IV. METHODOLOGY

The CIM model is intact by the application's technical details and includes both business process requirements and client needs.

BPMN is a standard for creating business processes. Application interface modeling is done using IFML, it comes

with a suite of tools for modeling user interaction and computer program behavior graphically.

In our methodology, we used the Eclipse tool to transform the BPMN model (CIM level) into the IFML model (PIM level) according to the MDA approach.

First of all, we created the BPMN and IFML metamodels shown in Fig. 2, Fig. 3, and Fig. 4 in Eclipse tool, then the creation of the BPMN model from the BPMN business process as a ".xmi" file from the BPMN metamodel, and we finally obtain the IFML model in ".xmi" file too by applying the transformation rules shown in Fig. 1 written in ATL language [5].

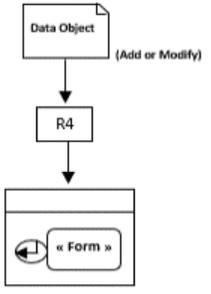
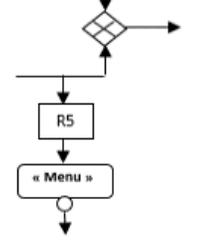
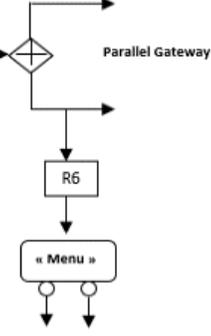
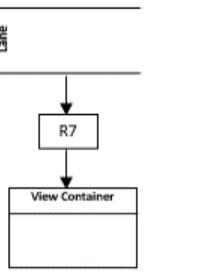
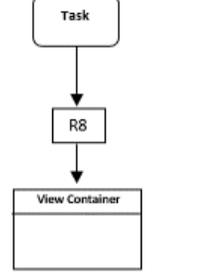
Lastly, we import the xmi file obtained to the commercial tool Webratio to generate the user interfaces automatically at the PSM level.

#### A. The CIM Level to the Pim Level Conversion Rules

BPMN models to IFML models transformation rules are shown in Fig. 1 every converting rule is defined in ATL and human language and represented by a schema.

Firstly, it is necessary to define the PIM level metamodel and the CIM level metamodel to describe ATL's transformation rules.

Human Language	Graphical representation	Transformation rule in ATL
-Each "State" of "Data Object" becomes "View Component"		<pre>Rule R 1{   from   stt: MMbpmn!StateObject   (stt.isTransformableStateObject ())   to   vcp : MMifml!ViewComponent{     Name &lt;- stt.name   } }</pre>
-Each "Display" state of "Data Object" becomes "View Container" with "List" view component plus "On Select Event"		<pre>Rule R 2{   from   stt : MMbpmn!StateObject   (stt.isTransformableStateObject () and   stt.isDisplay())   to   cn : MMifml!Container {     Name &lt;- '«WebPage»'+stt.name)   vcp : MMifml!ViewComponent(name&lt;-   '«List»',   ContainerBelongs&lt;-cn.name)   Osev : MMifml!OnSelectEvent{     ViewComponentBelongs&lt;-vcp.name)   } }</pre>
-Each "Delete" state of "Data Object" becomes " View Container" with "Delete" action plus simple event		<pre>Rule R 3{   from   stt : MMbpmn!StateObject   (stt.isTransformableStateObject () and   stt.isDelete())   to   cn : MMifml!Container {     Name &lt;- '«WebPage»'+stt.name)   ac : MMifml!Action (name&lt;- 'Delete',   ContainerBelongs&lt;-cn.name)   sev : MMifml!SimpleEvent(   ViewComponentBelongs&lt;-ac.name)   } }</pre>

<p>-Each "Add or Modify" state of "Data Object" becomes "View Container" with "Form" view component plus "On submit event"</p>		<pre> Rule R 4{   from   stt : MMBpmn!StateObject   (stt.isTransformableStateObject () and   (stt.isAdd() or stt.isModify()))   to   cn : MMifm!Container (   name &lt;- '&lt;WebPage&gt;'+stt.name)   vcp : MMifm!ViewComponent(   name&lt;-'Form'+ stt.name,   containerBelongs&lt;-cn.name)   osev : MMIFML!OnSubmitEvent(   ViewComponentBelongs&lt;-ac.name)   } </pre>
<p>-Each "Exclusive Gateway" becomes "Menu" view component with simple event corresponding transition in output "Exclusive Gateway"</p>		<pre> Rule R 5{   from   exl : MMBpmn!Exclusive   (exl.isTransformableEclusive ())   to   vcp: MMIFML!ViewComponent (   name&lt;- '&lt; Menu &gt;'+exl.name })   Rule R {   from   fl : MMBPMN!Flow   (fl.isTransformableFlow() and   fl.OutputisDecisionState())   To   MMIFML!SimpleEvent(name&lt;-fl.name,   ViewComponentBelongs&lt;-   Fl.NameDecisionStateOutput() )   } </pre>
<p>-Each "Parallel Gateway" becomes "Menu" view component with "Events" corresponding transition in output "Parallel Gateway"</p>		<pre> Rule R 6 {   from   prl : MMBpmn!Parallel   (prl.isTransformableParallel ())   to   vcp: MMIFML!ViewComponent (   name&lt;- '&lt; Menu &gt;'+prl.name })   Rule R {   from   fl : MMBPMN!Flow   (fl.isTransformableFlow() and   fl.OutputisDecisionState())   To   MMIFML!SimpleEvent(name&lt;-fl.name,   ViewComponentBelongs&lt;-   Fl.NameDecisionStateOutput() )   } </pre>
<p>-Each "Lane" becomes "View Container"</p>		<pre> Rule R 7{   from   ln : MMBpmn!Lane   (ln.isTransformableLane())   to   vcn : MMifm!ViewContainer (   Name &lt;- ln.name   )   } </pre>
<p>-Each "Task" becomes "View Component"</p>		<pre> Rule R 8{   from   tsk : MMBpmn!Task   ((not tsk.isManual()) and   (tsk.isTransformableTask()))   to   vcn : MMifm!ViewContainer (   Name &lt;- tsk.name   )   } </pre>

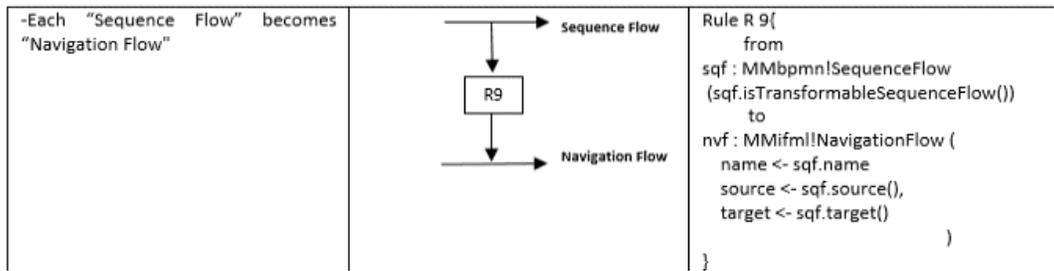


Fig. 1. CIM level to the PIM level transformation rules.

**B. CIM Level: BPMN Metamodel**

At the CIM level, BPMN was utilized to describe the business process because business processes are computationally independent, the OMG standard for business process modeling is named BPMN. Fig. 2 displays the BPMN metamodel [18].

**C. PIM Level: IFML Metamodel**

IFML Metamodel that uses the UML metamodel's main data types provides a set of UML metaclasses as the foundation for IFML metaclasses and presumes that the IFML ContentModel is defined in UML [19], [20].

The IFML metamodel is illustrated in Fig. 3 and Fig. 4.

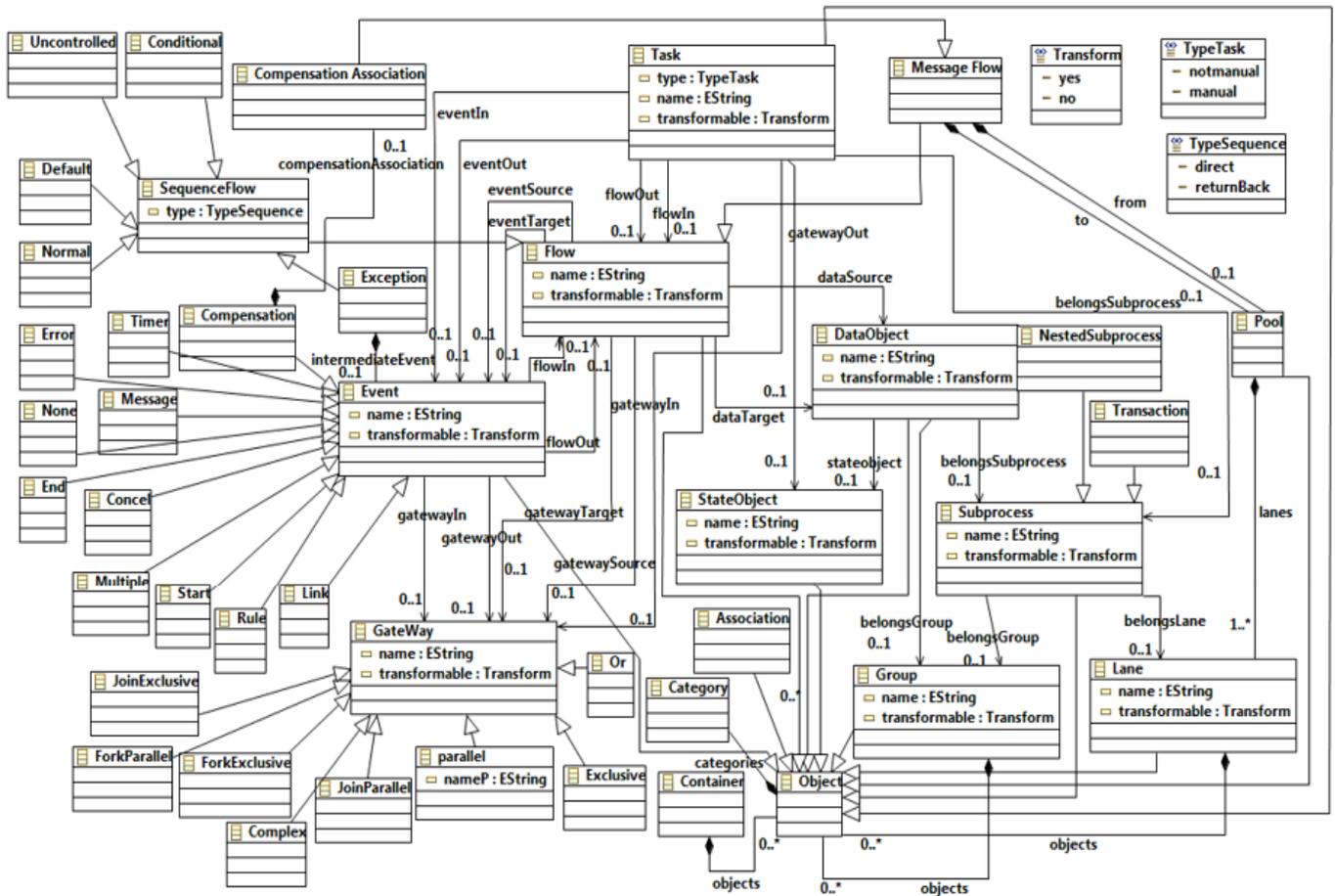


Fig. 2. The BPMN metamodel.

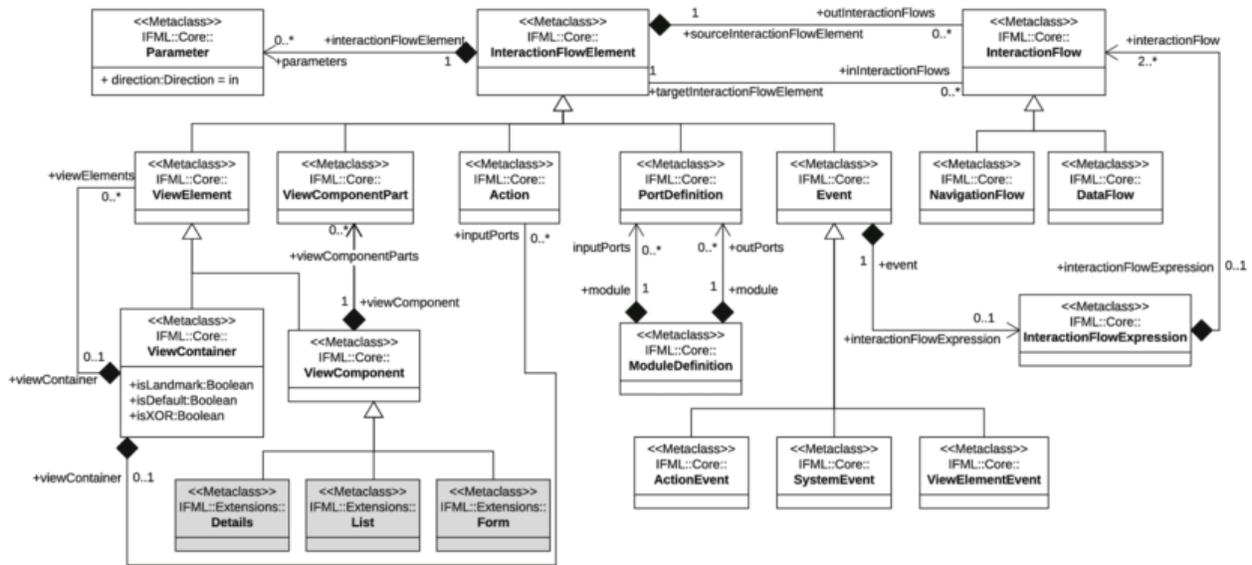


Fig. 3. IFML metamodel 1.

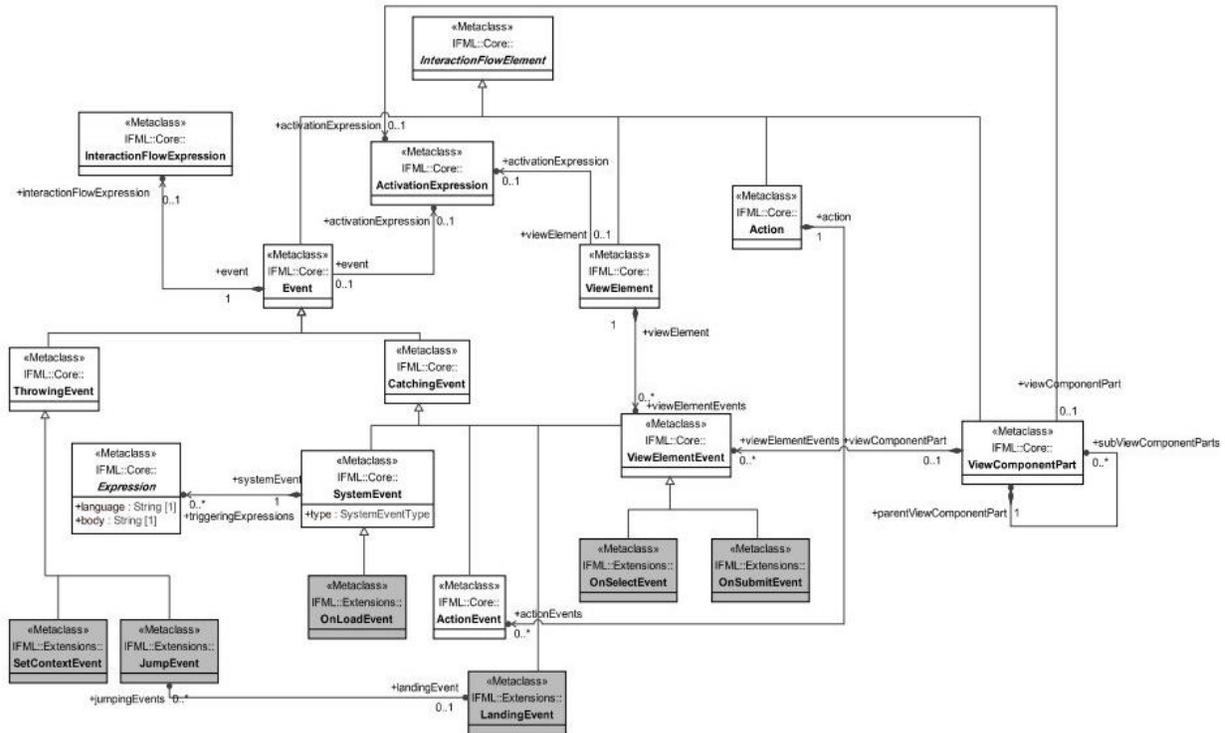


Fig. 4. IFML metamodel 2.

## V. CASE STUDY

The after-sales service is described, in which a customer can make CRUD operations (Create, Read, Update, and Delete) relating to a complaint if there is a problem regarding his order.

The process begins with authentication using a username and password. If the client is already registered on the

platform, he has access to a menu with four features, including add, modify, delete, and consult a complaint. If not, he must register to use the remaining features.

### A. CIM Level

The CIM level is illustrated by the BPMN diagram and BPMN business process, the first one is represented in Fig. 5, and the second is displayed in Fig. 6.

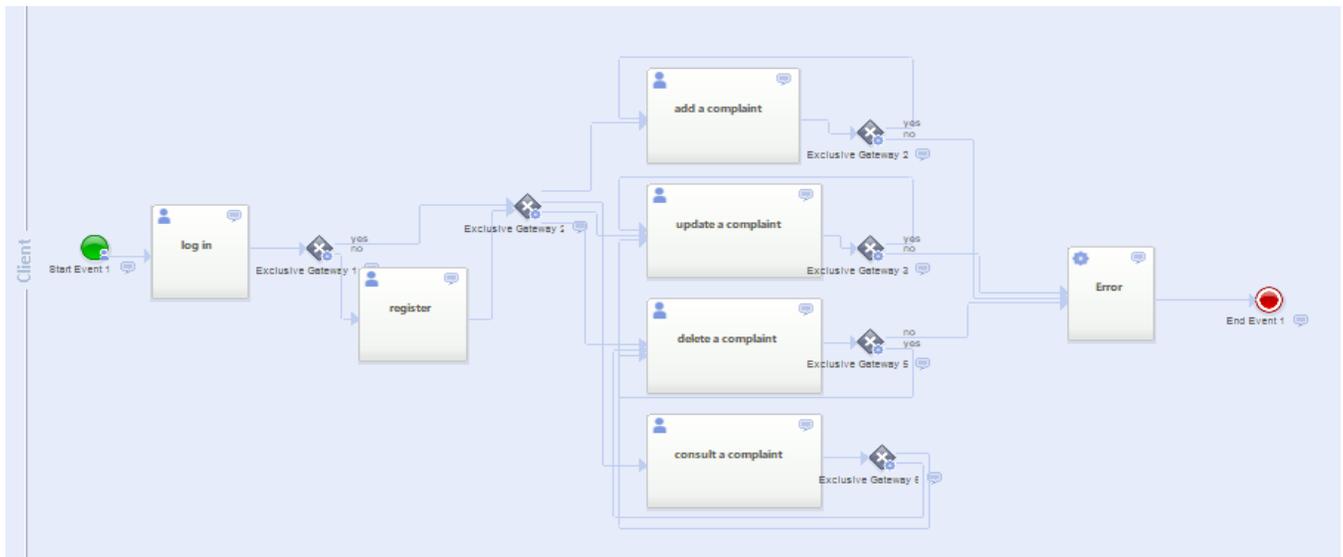


Fig. 5. BPMN diagram of the after-sales service.

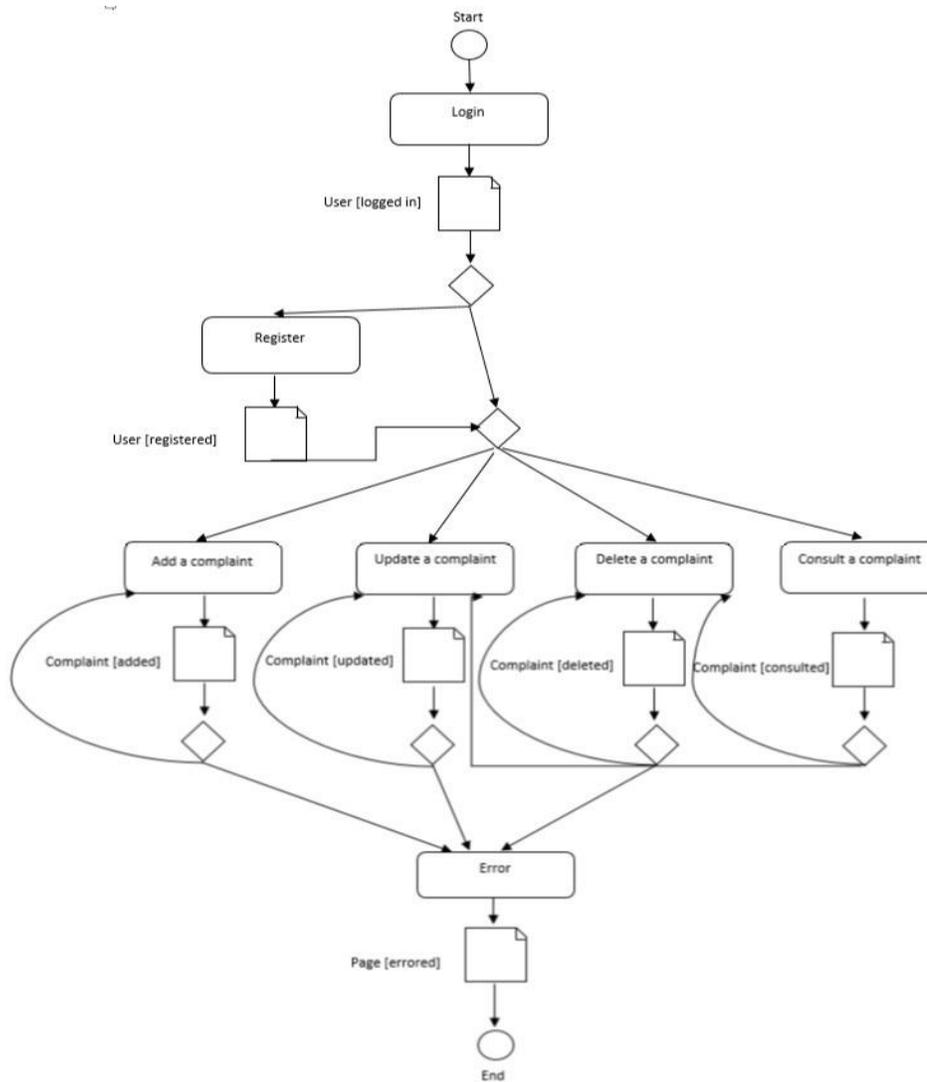


Fig. 6. BPMN business process diagram of the after-sales service.

### B. PIM Level

Webratio is the most powerful commercial tool available among others [8]. Some of the characteristics that motivated us to choose to work with this tool are:

- Provides graphical modeling; the IFML components are expanded further by this tool, so that users can design IFML diagrams for web applications.
- Based on the Eclipse and Webratio tools respond quickly.
- It is suitable for academic researchers.

We obtained the after-sales service front-end interfaces designed with the IFML language, by applying the shift rules from CIM to PIM. [5].

IFML development process involves modeling the application's domain first, then creating an IFML model using IFML language, and then converting the created model into a software system.

1) *Domain model*: A clear and accessible overview of information related to the application domain is provided by the domain model.

We don't suggest a novel modeling language for domain model modeling because a UML-based OMG standard is IFML; rather, we utilize UML class diagrams.

The case study's Domain model is described in Fig. 7 we added a table named complaint, User, Group, and Module are tables provided by default by the Webratio tool.

2) *IFML model*: In our case study, we include only two important elements in the example: the login/registration function and the client's space.

The public is the application's main page, allowing the user to authenticate himself to access the proposed features (Login page) or to register if it is his first connection to the application.

Otherwise, an error message is displayed, Fig. 8 shows the "Public" site view.

The site view Public contains an action definition (a description of a business logic that could be triggered by an event in an IFML model.) [21] "Register", which will allow the creation of a new user in the database and Fig. 9 shows its modeling.

The action definition Register begins with the input port, which receives the parameters from the user's registration form. The parameters are passed to the "Create User" operation which is based on the Entity "User."

We utilized a Data Flow to bind the information from the Input Port to the Output Port, the information between the input port and the corresponding "Create User", and the latter to the output port.

When these operations are completed successfully, the Action Definition must exit via the "Success" OK Port, otherwise, the "KO Port" handles the failed execution.

The complaint is the space dedicated to the user where he can perform one of the operations provided by CRUD (Create, Read, Update, or Delete a complaint).

In Fig. 10, the interface modeling is displayed.

In the site view Complaint, four areas were used to model CRUD operations. Each area is composed of pages and actions definitions, and also utilized view components (List, Form).

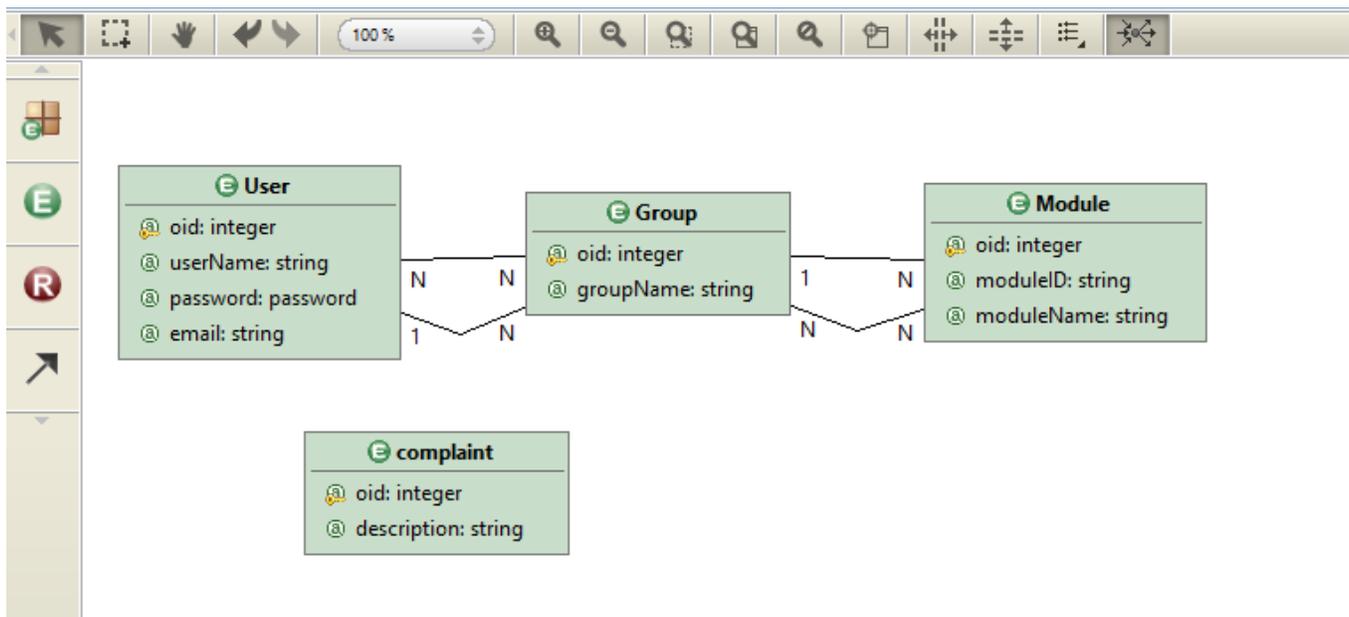


Fig. 7. Domain model of the after-sales service.

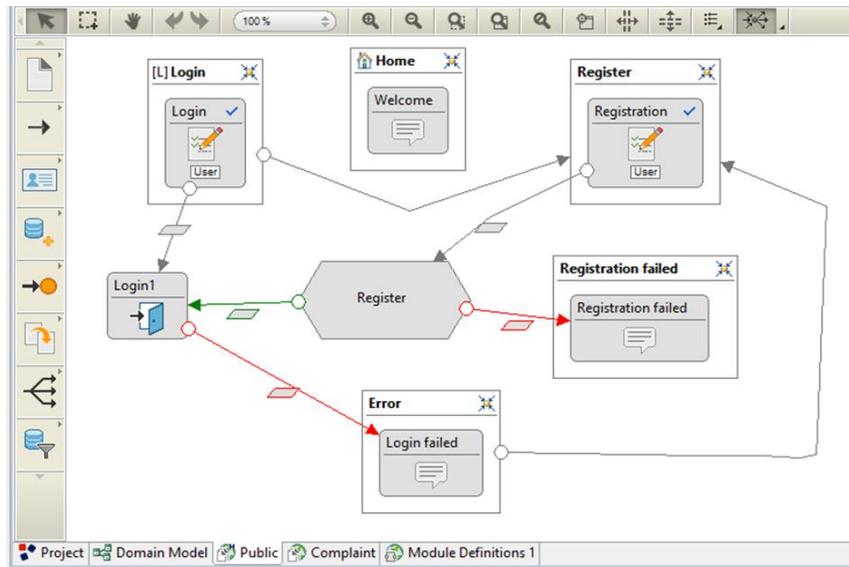


Fig. 8. Site view "Public".

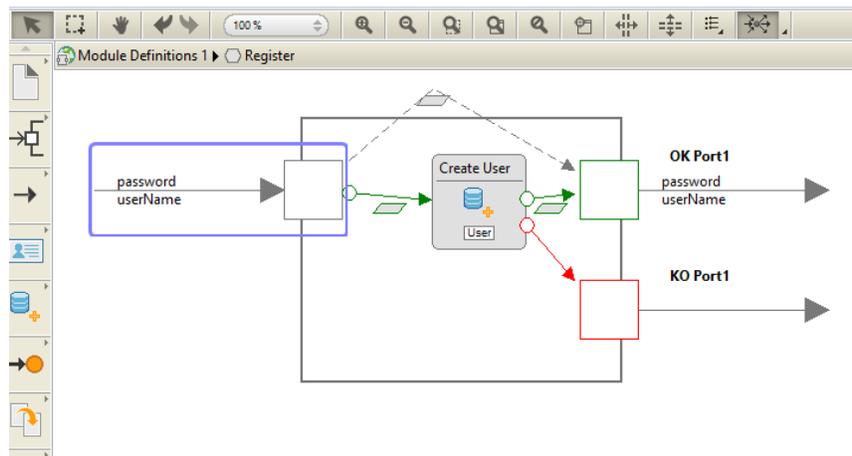


Fig. 9. Action definition "Register".

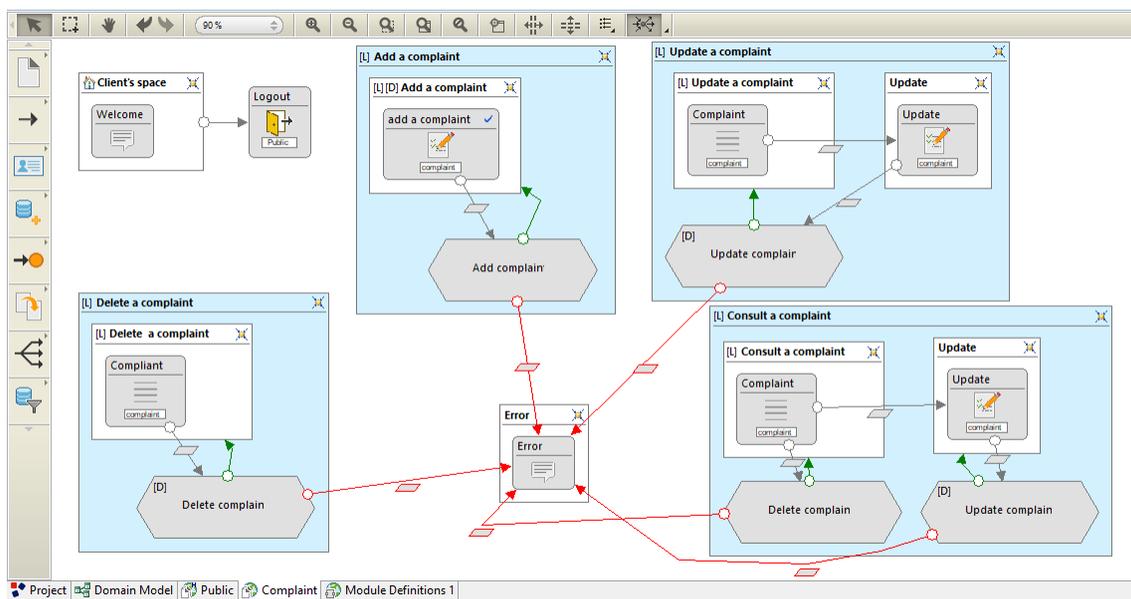


Fig. 10. Site view "Complaint".

To make the connection between the different pages and actions definitions we used flows (flow, OK flow, and KO flow) with parameters binding between source and target elements.

For example, to add a complaint we have a page named complaint containing a form. When the user presses the button “save”, the action definition Add a complaint is triggered, using a flow going from the form to the action definition, by editing parameters binding.

If the addition is done successfully we return to the Add a complaint page, otherwise we are redirected to an error page.

### C. PSM Level

In this paragraph, we will present some of the graphical user interfaces obtained by generating the application of the IFML model using the Webratio tool.

Fig. 11 shows the authentication page which provides the user to authenticate himself through a username and password.

Otherwise, he must register first by entering his personal information, and Fig. 12 displays this operation.

Another interface where the user can add a complaint is shown in Fig. 13.



Fig. 11. "Login" graphical interface.

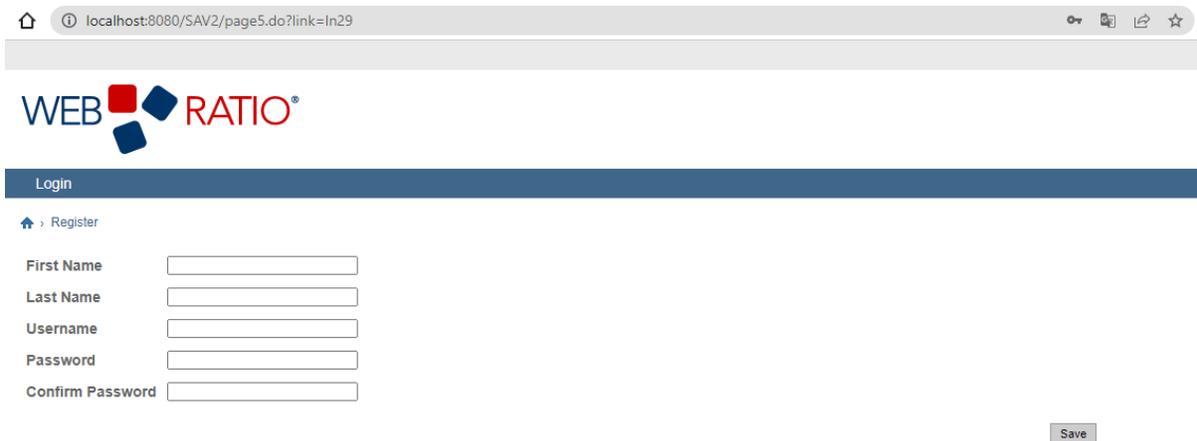


Fig. 12. "Register" graphical interface.

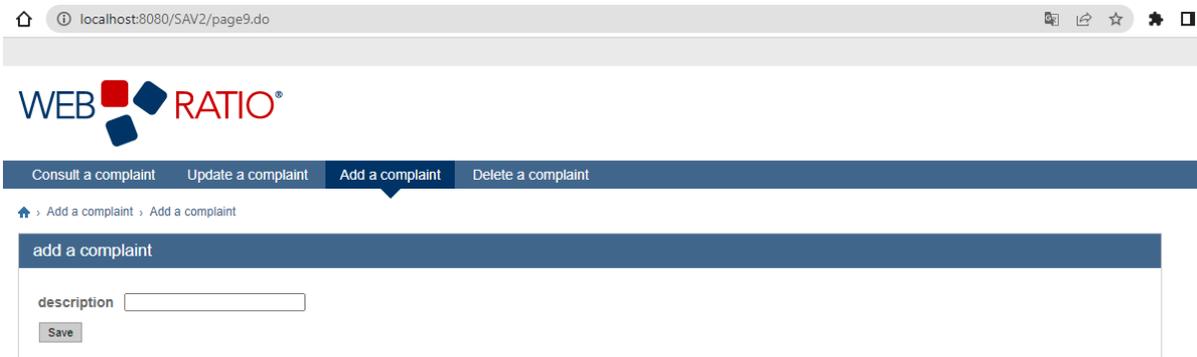


Fig. 13. "Add a complaint" graphical interface.

## VI. DISCUSSION AND LIMITATIONS

We may infer from an analysis of the works shown in Table I, that they can be divided into two groups. Those who were interested in the CIM to PIM shift and those who were interested in the PIM to PSM shift.

We can say that the majority of works use UML diagrams to present the different levels of the MDA approach, except [25] which used the BPMN standard that is utilized for business process modeling.

The UML class diagram is often utilized in most PIM to PSM transformations, IFML as a recommended standard for the modeling front-en interfaces at the PIM level is not widely used in the studied papers apart in [30].

The work [31] treated the model conversions in the MDA approach from the CIM level to the PSM via the PIM level, by using two models the first is for requirement and the second is a mathematical model represented by a graph.

The major advantage of our work which differs from other researches [33-40] is the fact that it covers all the model

transformations process in the MDA's approach. Starting from the CIM via the PIM level up to the PSM level, utilizing two standards BPMN and IFML, and a powerful commercial tool Webratio.

In our case study, we have limited it to CRUD operations, we can still extend it to cover other aspects of after-sales service, also the process of our methodology is semi-automatic in the importation of the xmi file in Webratio tool.

The CIM level to the PIM level shift is implemented in the Eclipse tool automatically, the IFML model is obtained in an xmi format [32] and imported into the Webratio tool to finally get the user interfaces.

The advantage of using IFML modeling is that the front-end interface of the program can be created quickly also the application is generated with a click.

But the technology has some disadvantages, for example, even though IFML modeling is still relatively a new technology, it is still best suited for modeling small programs rather than programs for large and medium-sized businesses.

TABLE I. COMPARISON OF PREVIOUS WORKS

Studied Papers	CIM to PIM shift		PIM to PSM shift	
	Reached	Illustration	Reached	Illustration
Rhazali et al. [22]	Yes	UML diagrams		
Kharmoum et al. [23]	Yes	UML class diagram		
Arrhioui et al. [24]	Yes	UML class diagram		
Khlif et al. [25]	Yes	BPMN		
Erraissi et al. [26]			Yes	UML class diagram
Deeba et al. [27]			Yes	UML class diagram
Srai et al. [28]			Yes	UML class diagram
Betari et al. [29]			Yes	UML class diagram
Koren et al. [30]			Yes	IFML
Zhang et al. [31]	Yes	Business-oriented Model Conceptual	Yes	Graph
Our methodology	Yes	BPMN	Yes	IFML

## VII. CONCLUSION

Developing software is easier with MDA's approach because it generates applications automatically from models in a shorter period and with absolutely no source code involvement.

Typically, web applications are built using Webratio's tool robust architecture, it is a consolidated industrial reality that enables any model to be created. In this sense, we suggested this work which highlighted the different model transformations.

Our methodology adheres to the MDA principles and suggests defining the various MDA levels' characteristics, as well as the CIM level to the PIM level shift rules utilizing two standards BPMN and IFML, till automatically generating the user interfaces utilizing the Webratio tool.

This is done by creating the two metamodels BPMN and IFML in Eclipse tool, after we elaborated the BPMN model of the case study then, by applying the shift rules in ATL language, we obtain the IFML model under .xmi format.

Finally, we imported this file in Webratio tool to get the graphical user interfaces.

The objective of improvement is the creation of a tool that will allow the model transformations from the CIM level until the obtention of the graphical user interfaces that will automatize the whole process of model transformations.

## REFERENCES

- [1] OMG, MDA Guide Version 2.0., 2014, [Online]. Available at: <http://www.omg.org/mda>.
- [2] M. Brambilla, A. Mauri and E. Umuhoza, "Extending the Interaction Flow Modeling Language (IFML) for Model Driven Development of Mobile Applications Front End", Awan, I., Younas, M., Franch, X., Quer, C. (eds) Mobile Web Information Systems. MobiWIS Lecture Notes in Computer Science, 2014, pp 176-191, vol 8640. Springer, Cham. DOI: 10.1007/978-3-319-10359-4\_15.
- [3] M. Brambilla, P. Fraternali. Interaction Flow Modeling Language: Model-Driven UI Engineering of Web and Mobile Apps with IFML (1st. ed.). Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2014, ISBN: 9780128001080, ISBN: 9780128005323. DOI:10.5555/2785641.
- [4] R. Acerbis, A. Bongio, M. Brambilla and S. Butti, Model-Driven Development Based on OMG's IFML with WebRatio Web and Mobile Platform. In: Cimiano, P., Frasinca, F., Houben, GJ., Schwabe, D. (eds)

- Engineering the Web in the Big Data Era. ICWE 2015. Lecture Notes in Computer Science(), 2015, vol 9114. Springer, Cham. DOI: 10.1007/978-3-319-19890-3\_39.
- [5] A. Sajji, Y. Rhazali and Y. Hadi, "A methodology for transforming BPMN to IFML into MDA", *Bulletin of Electrical Engineering and Informatics*, 11: 2773-2782, number 5,2022, pp. 2773-2782. DOI:10.11591/eei.v11i5.397.
- [6] S. Roubi , M. Erramdani and S. Mbarki, "A Model Driven Approach for generating Graphical User Interface for MVC Rich Internet Application", *Computer and Information Science*. 9. 10.5539/cis.v9n2p91. Press, Japan,2016, pp: 249-256. DOI: 10.1109/PNPM.1989.68558.
- [7] M. Brambilla , R. Cirami, M. Gasparini, A. Marassi and S. Pavanetto, "Code Generation based on IFML for the User Interfaces of the Square Kilometre Array (SKA)", *Proc. ICALEPCS'19*, New York, NY, USA, 2019, pp. 1307-1311. DOI: 10.18429/JACoW-ICALEPCS2019 WEPHA093.
- [8] N. Laaz, K. Wakil, S. Mbarki and D.N. Jawawi , "Comparative Analysis of Interaction Flow Modeling Language Tools". *Journal of Computer Science*, 14(9), 1267-1278, 2018. DOI: 10.3844/jcssp.2018.1267.1278.
- [9] K. Rong, X. Liu, " IFML-Based Web Application Modeling", *Procedia Computer Science*, 166, 2020 pp: 129- 133,ISSN: 1877-0509. DOI: 10.1016/j.procs.2020.02.034.
- [10] R. Rodriguez-Echeverria, C. Preciado, J. Sierra, M. Conejero , F. Sanchez-Figueroa,"AutoCRUD: Automatic generation of CRUD specifications in interaction flow modelling language", *Science of Computer Programming*,168, 2018, pp 165-168, ISSN 0167-6423 DOI: 10.1016/j.scico.2018.09.004.
- [11] M. Lachgar, A. Abdali, " Generating Android graphical user interfaces using an MDA approach", *Third IEEE International Colloquium in Information Science and Technology (CIST)*, 2014, pp. 80-85, DOI: 10.1109/CIST.2014.7016598.
- [12] E. Planas, G. Daniel, M. Brambilla, "Towards a model-driven approach for multi-experience AI-based user interfaces", *Softw Syst Model* 20, pp. 997–1009, 2021. DOI: 10.1007/s10270-021-00904-y.
- [13] M. Dumas, M. La Rosa, J. Mendling, A. Hajo, H. Reijers, *Fundamentals of Business Process Management*, London, 2018, ISBN: 978-3-662-56508-7. DOI 10.1007/978-3-642-33143-5.
- [14] OMG, *Business Process Model, and Notation (BPMN)*, 2013, [Online]. Available at: <https://www.omg.org/spec/BPMN/2.0/PDF>.
- [15] C. Bernaschina, M. Brambilla, A. Mauri, E. Umuhoza , "A Big Data Analysis Framework for Model-Based Web User Behavior Analytics", Cabot, J., De Virgilio, R., Torlone, R. (eds) *Web Engineering*, *Lecture Notes in Computer Science()*, ICWE ,2017, vol 10360. Springer, Cham. DOI: 0.1007/978-3-319-60131-16.
- [16] R. Acerbis et al., "WebRatio, an Innovative Technology for Web Application Development" ,Koch, N., Fraternali, P., Wirsing, M. (eds) *Web Engineering. ICWE. Lecture Notes in Computer Science*, vol 3140. Springer, Berlin, Heidelberg, 2004. DOI: 10.1007/978-3-540-27834-4\_90.
- [17] S. Ceri, P. Fraternali, A. Bongiotto, "Web Modeling Language (WebML): a modeling language for designing Web sites", *Computer Networks*, 33, 2000, pp. 137-157,ISSN :1389-1286. DOI: 10.1016/S1389-1286(00)00040-2.
- [18] Y. Rhazali, Y. Hadi, A. Mouloudi, "A Methodology of Model Transformation in MDA: from CIM to PIM", *International Review on Computers and Software (I.RE.CO.S.)*, vol. 10, no. 12, 2015, pp. 1186-1201. DOI: 10.15866/irecos.v10i12.8088.
- [19] M. Brambilla, E. Umuhoza, R. Acerbis, "Model-driven development of user interfaces for IoT systems via domain-specific components and patterns", *Internet Serv Appl* 8, 14,2017. DOI: 10.1186/s13174-017-0064-1.
- [20] OMG, *Interaction Flow Modeling Language*, 2015, [Online]. Available at: <https://www.omg.org/spec/IFML/1.0/Beta1/PDF>.
- [21] Acerbis, S., *Webratio* , 2001, [Online]. Available at: <https://my.webratio.com/learn/content?nav=65&link=oln208a.redirect&so=pcu1a>.
- [22] Y. Rhazali, Y. Hadi, I. Idriss Chana, M. Lahmer, A. Abdallah Rhattoy, "A Model Transformation in Model Driven Architecture from Business Model to Web Model", *IAENG International Journal of Computer Science*, 45:1, *IJCS\_45\_1\_16*, Vol. 45 Issue 1, pp104-117, vol. 45, no.1, 2018 [https://www.iaeng.org/IJCS/issues\\_v45/issue\\_1/IJCS\\_45\\_1\\_16.pdf](https://www.iaeng.org/IJCS/issues_v45/issue_1/IJCS_45_1_16.pdf).
- [23] N. Kharmoum, S. Ziti, Y. Rhazali, F. Omary, "An Automatic Transformation Method from the E3value Model to IFML Model: An MDA Approach", *Journal of Computer Science*, 15(6), 800-813, 2019. DOI: 10.3844/jcssp.2019.800.813.
- [24] K. Arrhioui, S. Mbarki , M. Erramdani, "Applying CIM-to-PIM Model Transformation for Development of Emotional Intelligence Tests Platform", *International Journal of Online and Biomedical Engineering (iJOE)*, 14(08), pp. 160–168,2018. DOI: 10.3991/ijoe.v14i08.8747.
- [25] W. Khlif , N.E. Ayed , H. Ben-Abdallah, "From a BPMN Model to an Aligned UML Analysis Model". *ICSOFT*, Porto, Portugal, 2018, pp. 623-631,. DOI: 10.5220/0006866606570665.
- [26] A. Erraissi, M. Banane, A. Belangour, M. Azzouazi," Big Data Storage using Model Driven Engineering: From Big Data Meta-model to Cloudera PSM meta-model", *International Conference on Data Analytics for Business and Industry: Way towards a Sustainable Economy (ICDABI)*, pp. 1-5,2020. DOI: 10.1109/ICDABI51230.2020.9325674.
- [27] F. Deeba, S. Kun, M. Shaikh, F. A. Dharejo, S. Hayat ,P. Suwansrikham, "Data transformation of UML diagram by using model driven architecture" ,*IEEE 3rd International Conference on Cloud Computing and Big Data Analysis (ICCCBDA)*,2018,pp. 300-303. DOI: 10.1109/ICCCBDA.2018.8386531.
- [28] A. Srail, A. F. Guerouate, N. Berbiche, H. HilalDrissi, "Generated PSM Web Model for E-learning Platform Respecting n-tiers Architecture", *International Journal of Emerging Technologies in Learning (IJET)*, 12(10), pp. 212–220, 2017. DOI: 10.3991/ijet.v12i10.7179.
- [29] O. Betari, M. Erramdani, S. Roubi, K. Arrhioui ,S. Mbarki, "Model Transformations in the MOF Meta-Modeling Architecture: From UML to CodeIgniter PHP Framework", Rocha, A. Serhini, M., Felgueiras, C. (eds) *Europe and MENA Cooperation Advances in Information and Communication Technologies. Advances in Intelligent Systems and Computing*, vol 520. Springer, Cham, 2017, pp. 227–234. DOI: 10.1007/978-3-319-46568-5\_23.
- [30] I. Koren, R. Klamma, "The Exploitation of OpenAPI Documentation for the Generation of Web Frontends", In *Companion Proceedings of the Web Crence 2018 (WWW '18)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 2018, pp. 781–787. DOI: 10.1145/3184558.3188740.
- [31] L. Zhang, H. Pingaud, E. Lamine, F. Fontanili, C. Bortolaso , M. Derras, "Model-Driven Approach to Transform Business Vision-Oriented Decision-Making Requirement into Solution-Oriented Optimization Model", Fujita, H., Fournier-Viger, P., Ali, M., Wang, Y. (eds) *Advances and Trends in Artificial Intelligence. Theory and Practices in Artificial Intelligence. IEA/AIE. Lecture Notes in Computer Science ()*, vol 13343. Springer, Cham, 2022, pp. 211–225. DOI: 10.1007/978-3-031-08530-7\_18.
- [32] OMG, "XML Metadata Interchange (XMI) Specification." 2015, [Online]. Available at: <https://www.omg.org/spec/XMI/2.5.1/PDF>.
- [33] A. Sajji, Y. Rhazali, Y. Hadi, "An Approach to Automate Generation of Graphical User Interfaces Through IFML", *Procedia Computer Science*, Vol 201, pp. 621-626, 2022, ISSN 1877-0509, DOI : 10.1016/j.procs.2022.03.081.
- [34] M. Ayadi, Y. Rhazali, M. Lahmer , "A Proposed Methodology to Automate the software manufacturing through Artificial Intelligence (AI)", *Procedia Computer Science*, Vol 201, pp. 627-631, 2022, ISSN 1877-0509 DOI: 10.1016/j.procs.2022.03.082.
- [35] S. Nasiri, Y.Rhazali, M. Lahmer, A. Adadi, "From User Stories to UML Diagrams Driven by Ontological and Production Model". *International Journal of Advanced Computer Science and Applications*, 12 (6), pp. 333 – 340, 2021 DOI: 10.14569/IJACSA.
- [36] N. Kharmoum, S. Ziti, Y. Rhazali,F. Omary, "A method of model transformation in mda approach from e3value model to bpmn2 diagrams in cim level". *IAENG International Journal of Computer Science*, vol. 46, no. 4, pp. 1–17, 2019. DOI: 10.47839/ijc.18.3.1524.
- [37] S. Nasiri, Y. Rhazali, A. Adadi, M. Lahmer, "Generation of User Interfaces and Code from User Stories". In: Joshi, A., Mahmud, M., Ragel, R.G. (eds), presented at *Information and Communication Technology for Competitive Strategies (ICTC 2021)*. *Lecture Notes in*

- Networks and Systems, vol 400, pp. 397-409, Springer, Singapore, ISBN 978-981-19-0094-5, 2023 DOI: 10.1007/978-981-19-0095-2\_38.
- [38] Y. Rhazali, Y. Hadi, A. Mouloudi, "Model Transformation with ATL into MDA from CIM to PIM Structured through MVC", *Procedia Computer Science*, 83, pp. 1096-1101, 2016. DOI:10.1016/j.procs.2016.04.229.
- [39] N. Kharmoum, S. Ziti, Y.Rhazali, F. Omary, "An automatic transformation method from the e3value model to uml2 sequence diagrams: An MDA approach". *International Journal of Computing*, vol. 18, no. 3, pp. 316–330, 2019 DOI:10.47839/ijc.18.3.1524.
- [40] Y. Rhazali, Y. Hadi, A. Mouloudi, (2016b), "CIM to PIM Transformation in MDA: from Service- Oriented Business Models to Web-Based Design Models", *International Journal of Software Engineering and Its Applications*. Vol. 10, no. 4, pp.125-142. DOI: 10.14257/ijseia.2016.10.4.13.

# A Fuzzy Logic based Solution for Network Traffic Problems in Migrating Parallel Crawlers

Mohammed Faizan Farooqui<sup>1</sup>, Mohammad Muqem<sup>2</sup>, Sultan Ahmad<sup>3\*</sup>, Jabeen Nazeer<sup>4</sup>, Hikmat A. M. Abdeljaber<sup>5</sup>

Department of Computer Application, Integral University, Lucknow, India<sup>1,2</sup>

Department of Computer Science-College of Computer Engineering and Sciences, Prince Sattam Bin Abdulaziz University, Al-Kharj, 11942, Saudi Arabia<sup>3\*,4</sup>

University Center for Research & Development (UCRD)-Department of Computer Science and Engineering, Chandigarh University, Gharuan, Mohali 140413, Punjab, India<sup>3</sup>

Department of Computer Science-Faculty of Information Technology, Applied Science Private University, Amman, Jordan<sup>5</sup>

**Abstract**—Search engines are the instruments for website navigation and search, because the Internet is big and has expanded greatly. By continuously downloading web pages for processing, search engines provide search facilities and maintain indices for web documents. Online crawling is the term for this process of downloading web pages. This paper proposes solution to network traffic problem in migrating parallel web crawler. The primary benefit of a parallel web crawler is that it does local analysis at the data's residence rather than inside the web search engine repository. As a result, network load and traffic are greatly reduced, which enhances the performance, efficacy, and efficiency of the crawling process. Another benefit of moving to a parallel crawler is that as the web gets bigger, it becomes important to parallelize crawling operations in order to retrieve web pages more quickly. A web crawler will produce pages of excellent quality. When the crawling process moves to a host or server with a specific domain, it begins downloading pages from that domain. Incremental crawling will maintain the quality of downloaded pages and keep the pages in the local database updated. Java is used to implement the crawler. The model that was put into practice supports all aspects of a three-tier, real-time architecture. An implementation of a parallel web crawler migration is shown in this paper. The method for efficient parallel web migration detects changes in the content and structure using neural network-based change detection techniques in parallel web migration. This will produce high-quality pages and detection for changes will always download new pages. Either of the following strategies is used to carry out the crawling process: either crawlers are given generous permission to speak with one another, or they are not given permission to communicate with one another at all. Both strategies increase network traffic. Here, a fuzzy logic-based system that predicts the load at a specific node and the path of network traffic is presented and implemented in MATLAB using the fuzzy logic toolbox.

**Keywords**—Web crawler; incremental crawling; fuzzy logic-based system; fuzzy logic toolbox

## I. INTRODUCTION

The Internet's numerous data sources, dynamic page production, and quick rate of change present a number of challenges for web crawling. All web crawlers are constructed using common components and must be scalable, reliable, and able to utilize available bandwidth effectively. When building a web crawler, politeness is a crucial issue that needs to be taken

into consideration [1]. Crawlers should refrain from overtaxing a web server by making numerous page requests quickly [2]. Web crawlers should adhere to the guidelines set forth by website administrators and should self-identify while seeking pages. The crawlers note a delay between two requests being sent to a web server simultaneously. Request Intervals are the name given to this waiting period [3]. Typically, 30 seconds elapses between downloads. A queue shuffling mechanism is used to enforce this waiting period; the queue is shuffled into a random order and distributed evenly among URLs coming from the same web server. The other crawler, similar to Mercator, implements their URL queue as a group of sub-queues, with a queue for each domain [4]. Each search engine keeps a central database of web pages on hand. In response to the query of user, search engine creates indexes for the repository. A web crawler, sometimes known as a spider, robot, or web pot, is a program that searches the Internet and collects web pages. Beginning with the seed URLs, the web crawler downloads the web document for those URLs. It takes new links from these downloaded documents and extracts them. Next, it is determined whether or not the extracted URLs have already been downloaded. URLs are redistributed to crawlers for additional downloading if it is confirmed that the documents have not already been downloaded. Up until there are no more URLs available for download, the process is repeated. Every day, a crawler downloads millions of online pages. The scheduler, multi-threaded downloader, queue for URLs and storage for text and metadata are all components of a web crawler. The search engine database must be updated often, and such updates must be incorporated by the web crawling system. To get over the processing bottleneck, multiple instances of this component operate simultaneously connected with very high bandwidth [5]. The HTTP protocol is used by web crawlers to download and process web pages from the internet. The downloader and the processor are the two components that make up the web crawling system. The processor receives the web pages that were downloaded by the downloader and processes them further. The HTTP protocol is used throughout this operation to download online pages. To download web pages, the downloader sends HTTP-GET queries. The pages are subsequently sent to the processor for additional processing and database integration. In order to download online pages from web servers, downloader adheres to a number of best practices. TCP Connection reusability,

\*Corresponding Author.

Compression, Conditional GETs, varying refresh frequency, and DNS caching are a few of the optimization strategies used on downloader to get around network constraints [6].

#### A. Reusability of TCP connections

To conserve time and network traffic, downloader maintain one TCP connection open for each IP address and utilize it repeatedly until the server's entire contents have been downloaded [7]. Web servers compress web pages before transmitting them to the requesting client in order to save network traffic [8]. Downloader use conditional GETs, which are defined in the HTTP protocol. Conditional GETs download the page if it changes after a certain date [9]. Different websites have different refresh rates; some are updated every few minutes, while others don't change for a year. More crucial than others are certain pages. Because of this, search engines sift through pages that are significant and likely to see frequent updates.

#### B. DNS Caching

DNS resolves and URL downloads are two factors that contribute to network congestion. Additional DNS server bottlenecks could occur, which would impact performance. DNS cache is used by the downloader or site crawling component to boost performance. When the size of the web expands, centralized web crawling techniques become inefficient. Only a very small portion of the web is crawled by search engines, and they frequently ignore significant pages. Because of this, simultaneous web crawlers are employed to reduce network and other bottlenecks.

The task of the processor is to gather, process, and store the downloaded web documents in the database of the search engine. It is responsible for collecting information from online documents and putting it in the database. The method involves taking keywords from a web page and ranking them. The database stores the frequency and position of keywords. In order to extract information, the processing component processes strings.

The remainder of this article is arranged accordingly. Section II reviews the related works. Section III gives a case study of crawler load, Section IV details about FIS and Fuzzy Logic. Section V gives the proposed solution, including the components. Section VI is having result discussion. Lastly, Section VII concludes the work.

## II. RELATED WORK

An illustration of a distributed publication system is the Web. Based on how they handle requests, the two models are centralized and distributed. In a distributed architecture, a central location generates the query, which is subsequently sent to a few distant locations for processing. The distant stations then relay their findings to the central location. The combined results are displayed to the searcher at a central location. Under contrast, one system in the centralized approach maintains the entire index there. The single central index is then used to test the searcher's query. When using a distributed design, the difficulty of searching each database is decreased because the index collection is spread across a number of databases. NetFind, WAIS, and Harvest are examples of the distributed

model. The authors of [10] introduced NetFind, a system that addressed the problem of finding information about people on the Web by employing a distributed search technique. WAIS employed a distributed set of indexes and a vector-space retrieval paradigm. WAIS struggled with its dearth of content.

According to the developers of [11], a Digest should only be used to convey queries to servers that have answers to such queries. Each server in the system creates a digest, which is then sent to a central site, containing a portion of the content that is available there. Archie, Veronica, and all search engines on the Internet are examples of the centralized paradigm. A number of existing extensively used documents (FTP files) were indexed by Archie [11]. The Gopher system could be searched thanks to Veronica [12]. Content from participating Gopher sites was downloaded and indexed at a single location. Publishing and accessing materials became simple thanks to the web. Additionally, it made following links easier [13]. The WWW Virtual Library is a method for finding resources on the Internet. A hierarchy was used to arrange the Virtual Library. Its quality was not exceptional because it was made by humans [14]. Three searchable indexes—the RBSE index, the WWW Worm, and Jumpstation—became well-known later in 1993. Each offered a searchable interface to a database of Web sites that was created automatically. Search engines like Lycos, InfoSeek, Yahoo, AltaVista, and Excite are a few examples. To fulfill a single searcher's request, the Meta search does numerous remote search engine queries. SavvySearch and MetaCrawler are two examples. These Meta engines use servers with content duplication. Because each search engine has a few unique documents in its index, meta-search is preferred. Additionally, as each search engine refreshes pages at a different rate, certain search engines will have current web content. Better results may be obtained via Meta search engines. Meta-search engines' disadvantage is a performance slowdown. AltaVista, Fast, and Google are a few examples of search engines.

A distributed system can scale by adding components, duplicate or distribute services so they are always available, and choose its components carefully to lower overall costs [15]. Three areas of distributed systems research are load balancing, resource allocation, and the overall design of distributed systems [16]. The goal of distributed system research is to create platforms that are consistent, dependable, and available. Locus, V Distributed System, and Eden are early distributed system models. Amoeba and Emerald are modern examples of distributed systems. The Coign system's creators claim that their system is capable of performing the kinds of optimizations that programmers often perform. Parallel searches are carried out simultaneously on all servers, integrated, ranked, and provided to the searcher. Ninja offers a platform for creating easily configurable, scalable Web services. The objective of Ninja is to reduce the administrative burden involved in managing a sizable cluster of systems [17]. Both distributed systems at the node level and parallel systems at the processor level require load balancing. The authors of [18] proposed a number of methods for assigning labour as well as detailed models for doing so. [19], which investigated load balancing in Amoeba, found that centralized decision-making outperforms distributed techniques. The Domain Name

System (DNS) has been used to disperse queries among several servers by having name servers answer with a list of addresses for a certain name. The authors of [20] suggested a mechanism for monitoring the load of the constituent systems.

In [21], the authors suggested DNS-based approaches that return server addresses in accordance with a model of the communication cost between the client and server. These methods address the issue of availability by limiting the addresses that name servers can distribute to servers that are known to be responsive. Very few significant pages are concealed within a big number of unimportant pages. How to get a good sample is the fundamental issue in web characterization? Pages containing scant or insignificant information should be excluded. It's important to gauge a website's importance. Web pages can be sampled using either vertical sampling or horizontal sampling. Web pages are gathered via vertical sampling, which is based on domain names. At many levels of the hierarchy, vertical sampling is possible. Vertical sampling at the highest level chooses nations with top-level domains like .in, .it, .au, etc. When vertical sampling is carried out at the second level, pages created by participants in the same institution or organization are collected (e.g. integraluniversity.ac.in).

A horizontal sample is the gathering of web pages based on selection criteria other than domain names. There are two ways to gather data: first, by employing a log of transactions in a major organization's proxy; and second, by using a web crawler. It is simple to locate popular pages when using a proxy, but the revisit duration cannot be adjusted because it depends on people. When using a web crawler, the popularity of the page must be approximated, but the revisit period may be adjusted.

### III. CASE STUDY OF CRAWLER LOAD

The quality of the data gathered during a crawl can always be enhanced. The sort of web graph search is determined by the ordering of the URL queue. By considering the pages' in-link factors, the queue can be sorted. The breadth first search can enhance the caliber of pages that are downloaded [22]. On the Internet, there are a lot of spam sites and indefinitely branching crawler traps whose pages are dynamically produced and made to have a very high in-link factor [23]. The various network metrics, including geographic distance and latency, are covered in this section.

#### A. Definition 1: Geographic Distance

On the Internet, there are resources that give a mapping between IP addresses and geographical data. Longitude and latitude are extracted from registrar address data by the existing Internet service [24]. Two hosts are controlled by the same ISP if their latitude and longitude are identical [25]. A pair of Internet hosts' latitude and longitude can be found, and their geographic separation can be determined by utilizing the spherical coordinates of the earth.

#### B. Definition 2: Latency

There are numerous methods for calculating the Round Trip Time between two Internet hosts. The UNIX Ping utility is used in the first technique, and the Trace route utility is used

in the second way [26]. The ICMP ECHO queries used by the Ping program are occasionally restricted or altered by ISPs. Some routers may prevent the TTL-restricted UDP packets that Trace route transmits [27].

#### C. Definition 3: Correlation between Metrics

Geographic distance and latency are strongly correlated [28]. The association between distance and RTT is stronger since the observations have lower linear distance values. A minimum end-to-end RTT is implied by linear distance along a path [29]. RTT and linear distance have a stronger correlation than RTT and end-to-end distance.

The client throughput in a conventional and active network was shown in Fig. 1. The X-axis indicates throughput, bits received by the clients at each simulated time, and Y-axis displays client's requests arrival rate [30]. Crawler throughput is proportional to client throughput for active indexing with 0% overhead. This establishes the comparability of the remaining samples. The throughput rapidly declines as the systems get saturated when both simulations reach a similar throughput of roughly 222 bits per tick. 140 bits per tick remain the consistent throughput at that point [30].

The throughput of a typical network crawler was shown in Fig. 2. The X-axis shows the overall arrival request rate, and the Y-axis shows the bits amount per simulated time unit that Web crawlers are receiving. The requests are started by both crawlers and actual customers. The typical client request delay for active indexing is shown in Fig. 3. The Y-axis indicates the delay in normal client response, and the X-axis reflects the rate at which requests are created by human clients. In traditional networks with 20 or 40 percent crawler traffic, the typical client delay is higher [30].

The relationship between request delay of crawler and the overall request arrival rate is seen in Fig. 4. As shown in the Fig. 3 and Fig. 4 the two curves are comparable, suggesting that the increased crawler load has no effect on the delay experienced by crawler sites [30].

The percentage of client requests that are always fulfilled is shown in Fig. 5. When the request arrival rate is low, all requests are nevertheless completed. The completion rates of customer requests have significantly decreased, as evidenced by the 20 percent and 40 percent crawler cases [30].

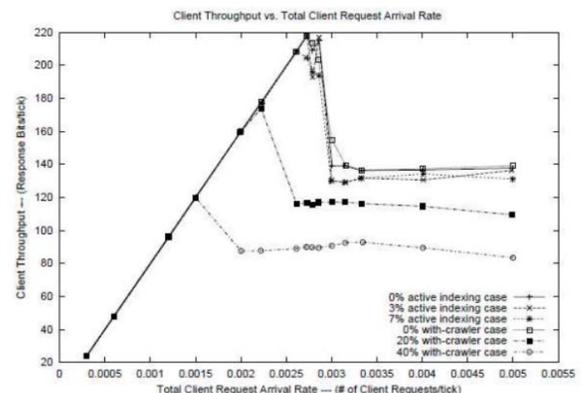


Fig. 1. Throughput of client in all cases [30].

#### IV. FIS AND FUZZY LOGIC

A FIS makes inferences from a knowledge base using a fuzzy inference engine (FIE). The FIE, represents the methods required for formalizing results and reasoning with the data in the knowledge database, is comparable to the brain of expert systems[31]. Boolean algebra's adaptation to deal with insufficient truth is fuzzy logic. Fuzzy logic illustrates the degree of truth of propositional logic. Everything in Boolean algebra may be stated in terms of binary numbers, or zero and one. Boolean algebraic values are replaced with the degree of truth in fuzzy logic.

The erroneous patterns of thinking are recorded using the level of truth. In an environment of ambiguity and uncertainty, this method of thinking is crucial to how humans make decisions. The membership function in fuzzy sets is comparable to the indicator function in classical set theory. Curves represent membership functions. According to member functions, every input space point is translated to a value lies between 0 and 1. A membership function has a triangular, bell-shaped, and trapezoidal shape. The universe of discourse is the name of the input space.

A fuzzy inference system is simpler to construct and has a very simple conceptual foundation. Three steps make up a fuzzy inference system: the input stage, the output stage, and the processing stage [32]. Input is translated into membership functions in the input step. At the processing stage, the appropriate rule is triggered, and each rule's result is generated before being combined. The output stage next transforms the outcome into output. The inference engine is the level of processing. The foundation of an inference engine is a set of IF-THEN logic rules. The THEN sub-statement is "consequent" if the IF sub-statement is "antecedent." A knowledge database contains n number of rules that are specific to fuzzy inference subsystems. The steps of the FIS are as follows:

- Fuzzification.
- Application of Fuzzy operators.
- Implication.
- Output Aggregation.
- Defuzzification.

Fuzzification of inputs is the process of assessing an input's degree of membership in its fuzzy sets using membership functions [33]. Fuzzy sets are used as the input and crisp values are produced during the defuzzification process. In fuzzy systems, there are two popular inference techniques. Ebrahim Mamdani proposed the first approach, the Mamdani fuzzy inference method, in 1975. Takagi-Sugeno-Kang proposed the second method, the Takagi-Sugeno-Kang fuzzy inference method, in 1985[34]. Numerous aspects of these approaches are comparable, such as the fuzzifying of the inputs and fuzzy operators. While Mamdani's inference uses fuzzy sets as its output membership functions. Sugeno's approach is computationally effective and functions well with adaptive and optimization strategies and uses output membership functions

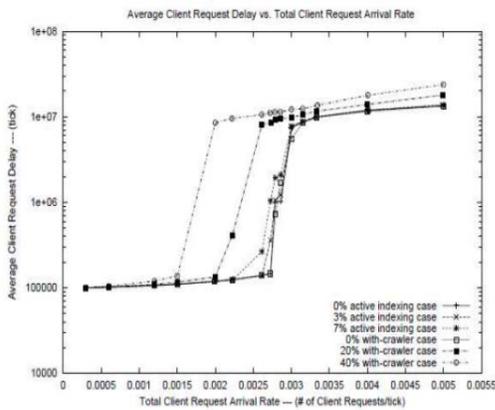


Fig. 2. Throughput of crawler.

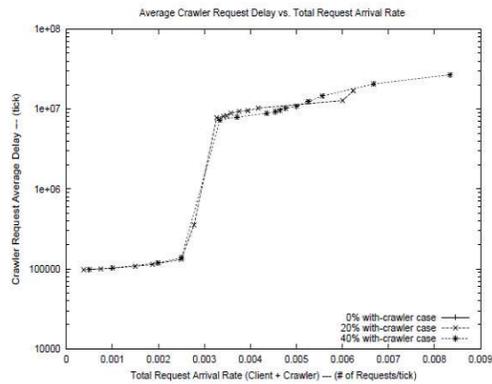


Fig. 3. Average client request delay in all cases [30].

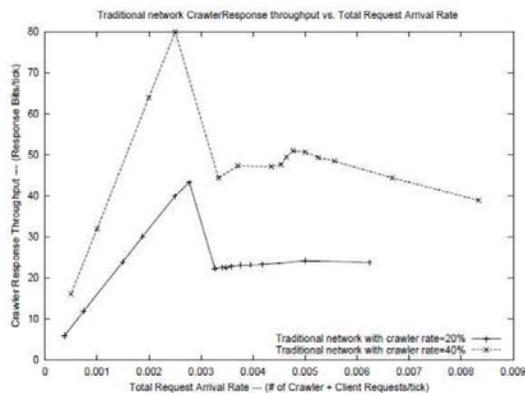


Fig. 4. Total request arrival time vs. average crawler request delay [30].

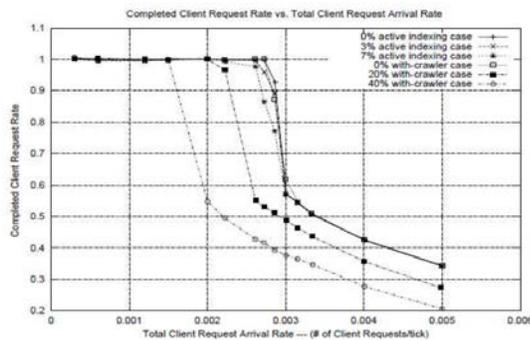


Fig. 5. Completed client request rates in all cases [30].

that are linear and constant. It also functions well when analysed mathematically.

The crawling procedure keeps the quality up. The web crawling is carried out using one of the two methods listed below: either the web crawlers are permitted to communicate with one another or they are not. Both methods add to the load on the network [35]. Here, a fuzzy logic-based method that predicts the load at a specific node and the path of network traffic is presented and implemented in MATLAB using the fuzzy logic toolbox.

### V. PROPOSED SOLUTION

These are the major steps of the proposed solution.

- a) Using a Fuzzy Inference System to Fix the Network Traffic Issue with Parallel Crawler Migration.
  - b) Using the membership function editor.
  - c) Using the Rule Editor to specify rules for a fuzzy inference system.
  - d) Rule Evaluation.
  - e) Adding up the results of the rule.
  - f) Removing fuzziness from the output value.
- 1) FIS to Solve Network Traffic problem in migrating parallel Crawlers.

The fuzzy set is the foundation of fuzzy logic theory. Every point in the input space is assigned a membership value between 0 and 1 that is defined by the curve known as the membership function. A fuzzy set is one that lacks a definite crisp border. Fuzzy Logic Toolbox includes the following tools for creating and customizing fuzzy inference systems:

- a) Fuzzy Inference System (FIS) Editor.
- b) Membership Function Editor.
- c) Rule Editor.
- d) Rule Viewer.
- e) Surface Viewer.

The Mamdani method is employed since it is well-liked for gathering knowledge. It enables us to speak more humanely when describing the expertise [36].

2) Defining FIS variables and fuzzification of the input variables using membership function editor.

a) *gaussmf*: The built-in membership function for the Gaussian curve in the fuzzy toolbox is known as *gaussmf*.  $y = \text{gaussmf}(x, [\text{sig } c])$  [37] gives the syntax (Fig. 6). The fuzzy toolbox's symmetric Gaussian function is dependent on the two parameters and as stated by.

$$f(x; \sigma, c) = e^{-\frac{(x-c)^2}{2\sigma^2}}$$

For example if  $y = \text{gaussmf}(x, [2 \ 5])$ .

`plot(x,y).`

`xlabel('gaussmf, P=[2 5]').`

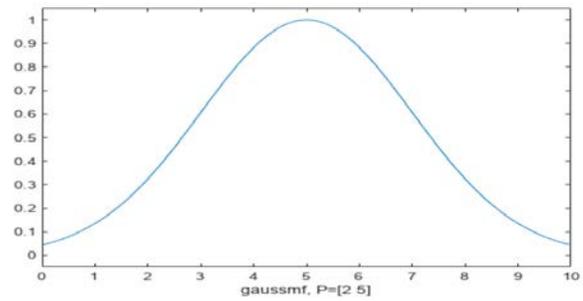


Fig. 6. Gaussmf Curve.

b) *trimf*: In the fuzzy toolbox, *trimf* is the built-in membership function with a triangular shape (Fig. 7). The triangular curve is a function of a vector  $x$  and depends on three parameters when the syntax is  $y = \text{trimf}(x, \text{params})$ ; if  $y = \text{trimf}(x, [a \ b \ c])$ :

$$f(x; a, b, c) = \begin{cases} 0, & x \leq a \\ \frac{x-a}{b-a}, & a \leq x \leq b \\ \frac{c-x}{c-b}, & b \leq x \leq c \\ 0, & c \leq x \end{cases}$$

Or,

$$f(x; a, b, c) = \max\left(\min\left(\frac{x-a}{b-a}, \frac{c-x}{c-b}\right), 0\right)$$

The triangle's base is indicated by first parameter  $a$  and third parameter  $c$ , while the triangle's peak is shown by second parameter  $b$  [38]. For example:

`x=0:0.1:10;`

`y=trimf(x,[3 6 8]);`

`plot(x,y)`

`xlabel('trimf, P=[3 6 8]')`

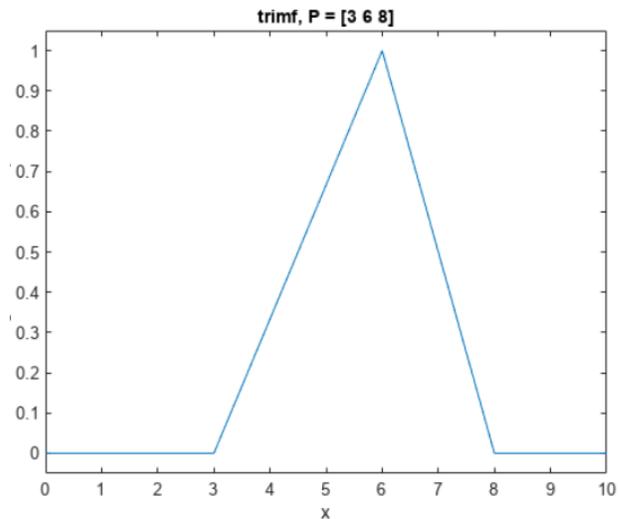


Fig. 7. Trimf function.

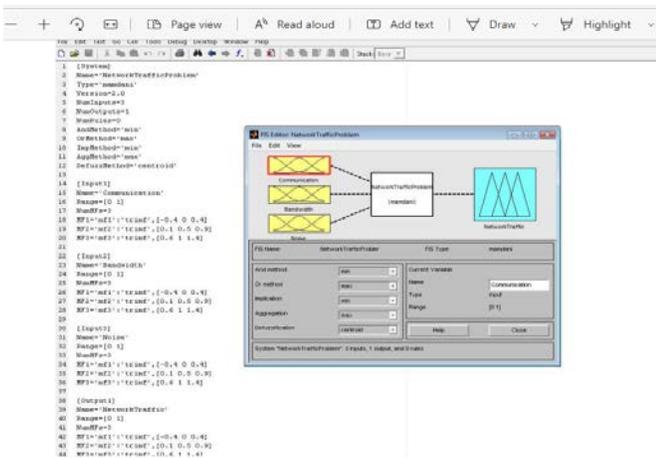


Fig. 8. FIS editor for network traffic problem.

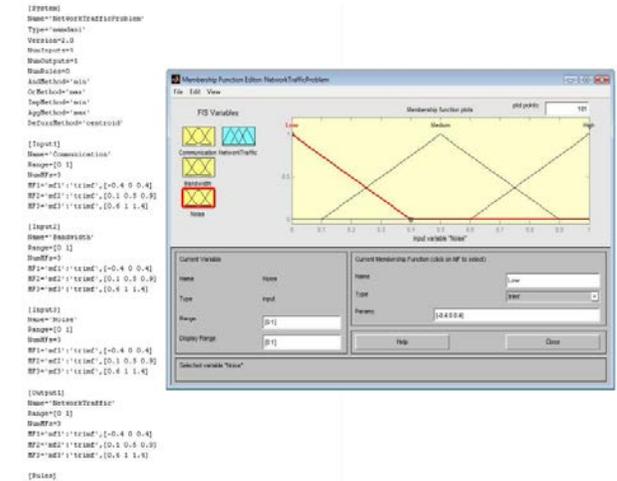


Fig. 11. FIS variable noise.

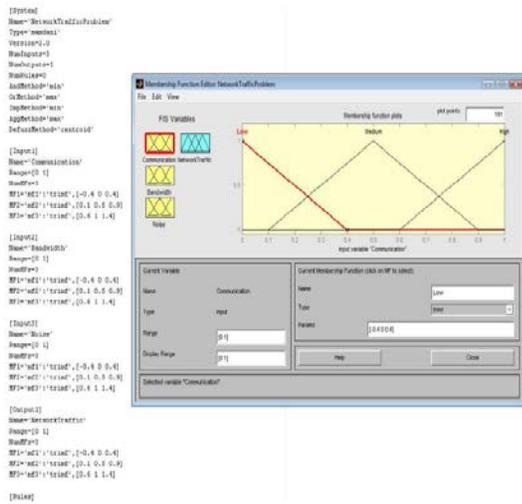


Fig. 9. FIS variable communication.

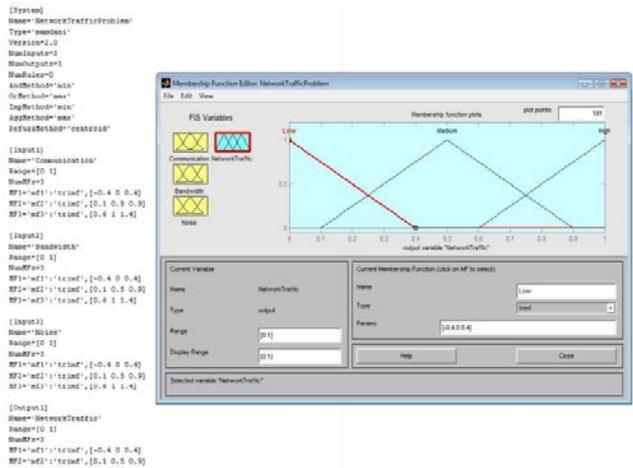


Fig. 12. FIS output variable network traffic.

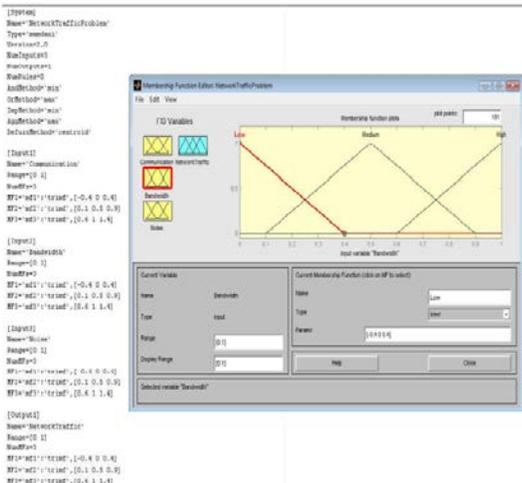


Fig. 10. FIS variable bandwidth.

The FIS editor for the Network Traffic Problem is shown in Fig. 8. The FIS variable Communication is shown in Fig. 9. The FIS variable Bandwidth is shown as a Fig. 10. The FIS variable Noise is shown in Fig. 11. The FIS output variable Network Traffic is shown in Fig. 12.

**Lemma 1.** When  $n$  is the total number of network nodes, the technique has a worst-case time complexity of  $O(n)$ .

**Proof.** Steps 1 and 3 of the method take a fixed amount of time. In the worst case situation, a node may have  $m$  number of gateways within its communication range, hence Step 2 can be finished in  $O(m)$  time. The node selects a CH in Step 4 in  $O(n)$  time and  $O(m)$  processing time after resetting its backup set (Step 4.2) in Step 4. (Step 4). The worst-case execution time for Step 4 is  $O(m) + O(n)$ , or  $O(n)$ , if  $n > m$ . (Step 4). Step 4 shows that the worst-case processing time of the algorithm is  $O(n)$ .

**Lemma 2.** The algorithm's worst-case message exchange complexity is  $O(1)$  per node or  $O(n)$  across the network's  $n$  nodes.

**Proof.** During the cluster creation phase, a node calculates the cost values of the CHs within its communication range and

sends a join request to the selected CH. Nodes that cannot communicate with any CHs, however, broadcast a HELP request message (step 4.1). It will send a join request message to join the cluster using multi-hop communication if it finds a node that can help. Therefore, in the worst scenario, a node only needs to send two messages for the cluster to form. As a result, each node's message complexity,  $O$ , is constant (1). The network's overall message exchange complexity is  $O$  as a result (n).

3) Using the Rule Editor to specify rules for a fuzzy inference system to solve the network traffic issue when migrating parallel crawlers.

H	L	M	M
H	L	H	M
H	M	L	M
H	M	M	M
H	M	H	H
H	H	L	M
H	H	M	H
H	H	H	H

4) Rule evaluation, rule output aggregate, and output value defuzzification.

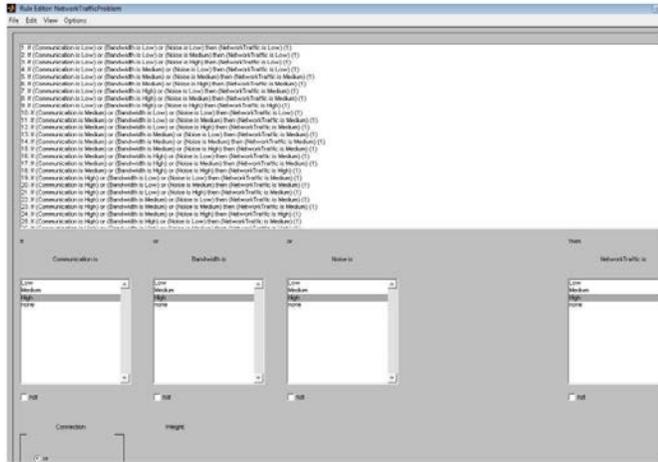


Fig. 13. Rules editor for network traffic problem.

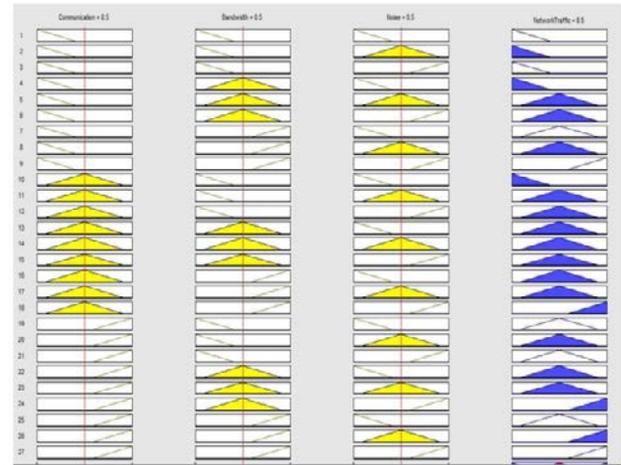


Fig. 14. Rule evaluation aggregation of the rule output.

TABLE I. RULES IN FIS (WHERE L=LOW, M=MEDIUM, H=HIGH)

Communication	Bandwidth	Noise Network	Traffic
L	L	L	L
L	L	M	L
L	L	H	L
L	M	L	L
L	M	M	M
L	M	H	M
L	H	L	M
L	H	M	M
L	H	H	H
M	L	L	L
M	L	M	M
M	L	H	M
M	M	L	M
M	M	M	M
M	M	H	M
M	H	L	M
M	H	M	M
M	H	H	H
M	H	M	M
M	H	H	H
H	L	L	M

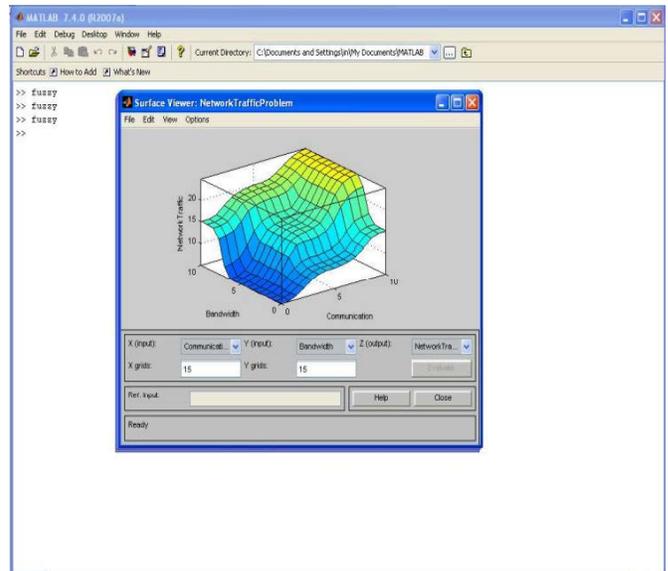


Fig. 15. Surface viewer for network traffic problem.

The FIS rules are in Table I. The Rules Editor for Network Traffic Problem is shown in Fig. 13. The Rule Evaluation Aggregation of the rule output is shown in Fig. 14. The Surface Viewer for Network Traffic Problem is shown in Fig. 15.

VI. RESULT AND DISCUSSION

The algorithm is integrated with the aforementioned module. The MATLAB Compiler is used to generate the code. The Implementation is tested against current web crawlers and designed to function on active websites.

TABLE II. LOAD CAUSED USING CONVENTIONAL CRAWLER

	Page 1	Page 2	Page 3	Total Load (KB)
visit 1	185	185	185	555
visit 2	193	196	195	
visit 3	188	189	199	
visit 4	200	201	205	
visit 5	188	199	188	
load caused	954	970	972	2896
visit 6	188	189	188	
visit 7	198	198	189	
visit 8	178	176	189	
visit 9	189	187	189	
visit 10	199	189	198	
load caused	1906	1906	1925	5740

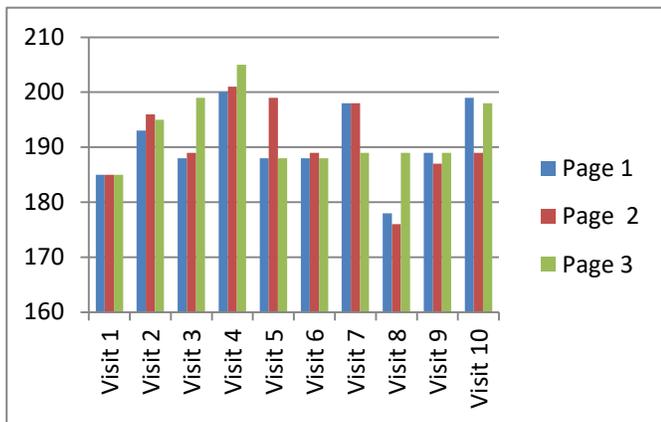


Fig. 16. Load caused using conventional crawler.

TABLE III. LOAD CAUSED USING SINGLE THREADED CRAWLER

	Page 1	Page 2	Page 3	Total Load (KB)
visit 1	78	87	98	263
visit 2	87	89	98	
visit 3	76	98	98	
visit 4	87	98	87	
visit 5	87	998	89	
load caused	415	470	470	1355
visit 6	87	89	87	
visit 7	78	98	98	
visit 8	98	76	98	
visit 9	87	97	98	
visit 10	78	98	87	
load caused	843	928	938	2709

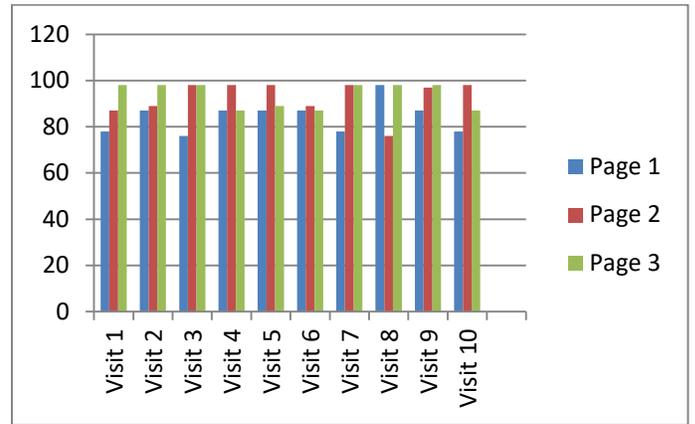


Fig. 17. Load caused using Single threaded crawler.

TABLE IV. LOAD CAUSED USING AGENT BASED CRAWLER

	Page 1	Page 2	Page 3	Total Load (KB)
visit 1	35	35	37	107
visit 2	36	37	37	
visit 3	43	36	45	
visit 4	34	45	57	
visit 5	34	43	43	
load caused	182	196	219	597
visit 6	43	53	43	
visit 7	43	34	34	
visit 8	45	54	43	
visit 9	34	43	45	
visit 10	34	34	45	
load caused	381	414	429	1224

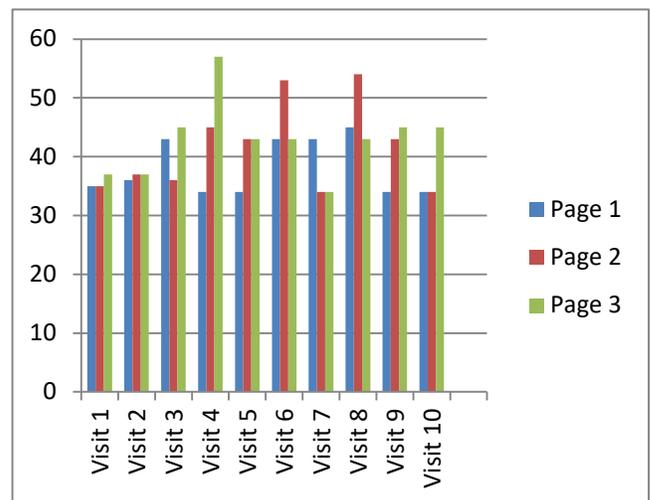


Fig. 18. Load caused using agent based crawler.

The load produced by a conventional crawler is shown in Table II. The load produced by a single threaded crawler is shown in Table III. The load produced by an agent-based crawler is seen in Table IV. The load created by migrating parallel web crawlers is shown in Table V. The graph in Fig.

16 to Fig. 19 depict the network load created by various methods. Three websites are used in the analysis and comparison of the methods. Since an HTML page typically weighed 205 KB, the network traffic produced by the conventional centralised crawling strategy was 555 KB. While in our method, the pages were compressed on the server, and the traffic load that was discovered was 70 KB. As seen in the above Fig. 20, the load incurred after five visits to the pages was 2896 KB, 1355 KB, 597 KB, and 379 KB, respectively, and after 10 visits, the load was 5740 KB, 2709 KB, 1224 KB, and 774 KB, respectively. Additionally, this resulted in less network traffic.

TABLE V. LOAD CAUSED USING MIGRATING PARALLEL WEB CRAWLER

	Page 1	Page 2	Page 3	Total Load (KB)
visit 1	23	23	24	70
visit 2	24	24	24	
visit 3	24	28	27	
visit 4	27	26	27	
visit 5	24	27	27	
load caused	122	128	129	379
visit 6	27	27	27	
visit 7	26	26	26	
visit 8	26	26	27	
visit 9	27	25	27	
visit 10	25	26	29	
load caused	253	258	263	774

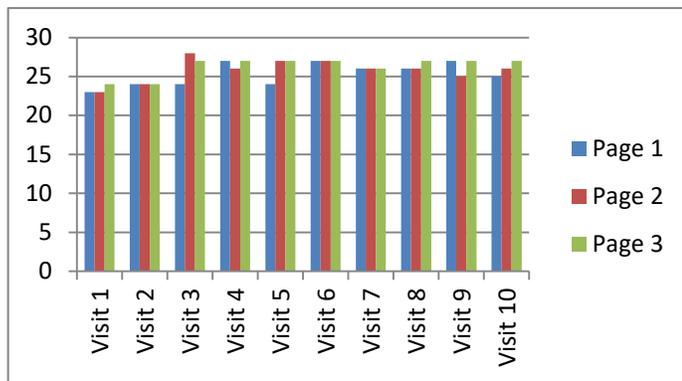


Fig. 19. Load caused using migrating parallel web crawler.

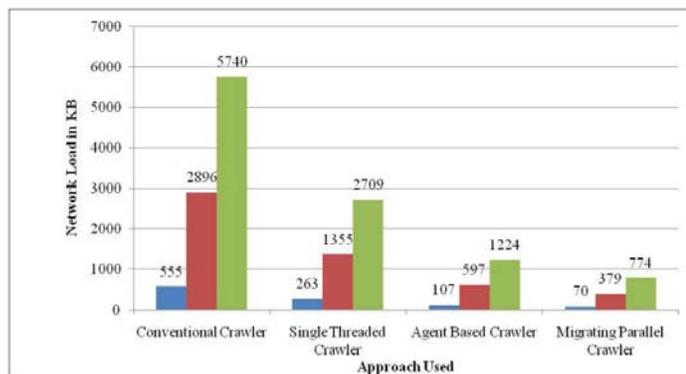


Fig. 20. Graph showing network load caused in various approaches.

## VII. CONCLUSION

This paper discusses the crawling process using one of the two following approaches: either allowing crawlers to communicate among themselves freely or forbidding them from doing so altogether. Both approaches increase network traffic. Here, a fuzzy logic-based method that predicts the load at a specific node and the path of network traffic is presented and implemented in MATLAB using the fuzzy logic toolbox. The experimental findings demonstrate that the network demand is decreased when a parallel web crawler is migrated.

## ACKNOWLEDGMENT

We thank the Deanship of Scientific Research, Prince Sattam Bin Abdulaziz University, Alkharj, Saudi Arabia for help and support.

## REFERENCES

- [1] M. Haleem, M. F. Farooqui, and M. Faisal, "Tackling Requirements Uncertainty in Software Projects: A Cognitive Approach," *Int. J. Cogn. Comput. Eng.*, vol. 2, pp. 180–190, 2021.
- [2] S. Khalil and M. Fakir, "RCrawler: An R package for parallel web crawling and scraping," *SoftwareX*, vol. 6, pp. 98–106, 2017.
- [3] Jha S, Sultan A, Alharbi M, Alouffi B, Sebastian S. Secured and provisioned access authentication using subscribed user identity in federated clouds. *International Journal of Advanced Computer Science and Applications*. 2021;12(11).
- [4] S. Singh and N. Tyjagi, "A Novel Architecture of Mercator: A Scalable, Extensible Web Crawler with Focused Web Crawler," *Int. J. Comput. Sci. Mob. Comput.*, vol. 2, no. 6, pp. 244–250, 2013.
- [5] Gopi R, Mathapati M, Prasad B, Ahmad S, Al-Wesabi FN, Abdullah Alohalhi M, Mustafa Hilal A. Intelligent DoS attack detection with congestion control technique for vanets. *Computers, Materials & Continua*. 2022;72(1):141-56.
- [6] Ahmad S, Jha S, Alam A, Alharbi M, Nazeer J. Analysis of Intrusion Detection Approaches for Network Traffic Anomalies with Comparative Analysis on Botnets (2008–2020). *Security and Communication Networks*. 2022 May 12;2022.
- [7] J. Kim, H. Kim, and J. Rexford, "Analyzing traffic by domain name in the data plane," in *Proceedings of the ACM SIGCOMM Symposium on SDN Research (SOSR)*, 2021, pp. 1–12.
- [8] R. Palacios, A. F. Fernández-Portillo, E. F. Sánchez-Úbeda, and P. García-De-Zúñiga, "HTB: A Very Effective Method to Protect Web Servers Against BREACH Attack to HTTPS," *IEEE Access*, vol. 10, pp. 40381–40390, 2022.
- [9] R. P. Kasturi et al., "Mistrust Plugins You Must: A Large-Scale Study Of Malicious Plugins In WordPress Marketplaces," in *31st USENIX Security Symposium (USENIX Security 22)*, 2022, pp. 10–12.
- [10] M. F. Schwartz and C. Pu, "Applying an information gathering architecture to Netfind: a white pages tool for a changing and growing Internet," *IEEE/ACM Trans. Netw.*, vol. 2, no. 5, pp. 426–439, 1994.
- [11] M. Li et al., "Bringing Decentralized Search to Decentralized Services," in *15th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 21)*, 2021, pp. 331–347.
- [12] B. E. É. R. Oliveira, "Web Search Engines-A study on the evolution of user interfaces," 2021.
- [13] E. F. Pettersen et al., "UCSF ChimeraX: Structure visualization for researchers, educators, and developers," *Protein Sci.*, vol. 30, no. 1, pp. 70–82, 2021.
- [14] L. Heng, G. Yin, and X. Zhao, "Energy aware cloud - edge service placement approaches in the Internet of Things communications," *Int. J. Commun. Syst.*, vol. 35, no. 1, p. e4899, 2022.
- [15] Uddin M. Y, Ahmad S. A review on edge to cloud: paradigm shift from large data centers to small centers of data everywhere. In *2020 International Conference on Inventive Computation Technologies (ICICT) 2020 Feb 26 (pp. 318-322)*. IEEE.

- [16] I. K. Aksakalli, T. Çelik, A. B. Can, and B. Tekinerdoğan, "Deployment and communication patterns in microservice architectures: A systematic literature review," *J. Syst. Softw.*, vol. 180, p. 111014, 2021.
- [17] C. S. Long et al., "California Needs Clean Firm Power, and So Does the Rest of the World: Three Detailed Models of the Future of California's Power System all show that California needs Carbon-Free Electricity Sources that don't Depend on the Weather," *Clean Air Task Force*, 2021.
- [18] A. M. Fathollahi-Fard, L. Woodward, and O. Akhrif, "Sustainable distributed permutation flow-shop scheduling model based on a triple bottom line concept," *J. Ind. Inf. Integr.*, vol. 24, p. 100233, 2021.
- [19] K. Oshima, D. Yamamoto, A. Yumoto, S.-J. Kim, Y. Ito, and M. Hasegawa, "Online machine learning algorithms to optimize performances of complex wireless communication systems," *Math. Biosci. Eng.*, vol. 19, no. 2, pp. 2056–2094, 2022.
- [20] G. L. Golewski, "Evaluation of fracture processes under shear with the use of DIC technique in fly ash concrete and accurate measurement of crack path lengths with the use of a new crack tip tracking method," *Measurement*, vol. 181, p. 109632, 2021.
- [21] L. Csikor, H. Singh, M. S. Kang, and D. M. Divakaran, "Privacy of DNS-over-HTTPS: Requiem for a Dream?," in *2021 IEEE European Symposium on Security and Privacy (EuroS&P)*, 2021, pp. 252–271.
- [22] A. Garg, K. Gupta, and A. Singh, "Survey of Web Crawler Algorithms," *Int. J. Adv. Res. Comput. Sci.*, vol. 8, no. 5, 2017.
- [23] A. Goswami and A. Kumar, "Online Social Communities," *Digit. Bus.*, pp. 289–341, 2019.
- [24] M. Mansoori and I. Welch, "How do they find us? A study of geolocation tracking techniques of malicious web sites," *Comput. Secur.*, vol. 97, p. 101948, 2020.
- [25] X. Mi et al., "Resident evil: Understanding residential IP proxy as a dark service," in *2019 IEEE symposium on security and privacy (SP)*, 2019, pp. 1185–1201.
- [26] I. Pelle, T. Lévai, F. Németh, and A. Gulyás, "One tool to rule them all: A modular troubleshooting framework for SDN (and other) networks," in *Proceedings of the 1st ACM SIGCOMM Symposium on Software Defined Networking Research*, 2015, pp. 1–7.
- [27] I. D. Oladipo, M. AbdulRaheem, J. B. Awotunde, A. K. Bhoi, E. A. Adeniyi, and M. K. Abiodun, "Machine Learning and Deep Learning Algorithms for Smart Cities: A Start-of-the-Art Review," *IoT IoE Driven Smart Cities*, pp. 143–162, 2022.
- [28] M. Haleem, M. F. Farooqui, and M. Faisal, "Cognitive impact validation of requirement uncertainty in software project development," *International Journal of Cognitive Computing in Engineering*, vol. 2, pp. 1–11, 2021, doi: 10.1016/j.ijcce.2020.12.002.
- [29] J. Jiang et al., "Via: Improving internet telephony call quality using predictive relay selection," in *Proceedings of the 2016 ACM SIGCOMM Conference*, 2016, pp. 286–299.
- [30] X. Yuan, M. H. MacGregor, and J. Harms, "An efficient scheme to remove crawler traffic from the internet," in *Proceedings. Eleventh International Conference on Computer Communications and Networks*, 2002, pp. 90–95.
- [31] S. Sweta and K. Lal, "Personalized adaptive learner model in e-learning system using FCM and fuzzy inference system," *Int. J. Fuzzy Syst.*, vol. 19, no. 4, pp. 1249–1260, 2017.
- [32] H. Uğuz, "Adaptive neuro-fuzzy inference system for diagnosis of the heart valve diseases using wavelet transform with entropy," *Neural Comput. Appl.*, vol. 21, no. 7, pp. 1617–1628, 2012.
- [33] M. Taki and Y. Omid, "A new fuzzy based joint DF relay selection and link adaptation," in *2015 International Conference on Communications, Signal Processing, and their Applications (ICCSPA'15)*, 2015, pp. 1–6.
- [34] V. Jain and S. Raheja, "Improving the prediction rate of diabetes using fuzzy expert system," *IJ Inf. Technol. Comput. Sci.*, vol. 7, no. 10, pp. 84–91, 2015.
- [35] G. Sun, R. Liang, H. Qu, and Y. Wu, "Embedding spatio-temporal information into maps by route-zooming," *IEEE Trans. Vis. Comput. Graph.*, vol. 23, no. 5, pp. 1506–1519, 2016.
- [36] G. Improta, V. Mazzella, D. Vecchione, S. Santini, and M. Triassi, "Fuzzy logic-based clinical decision support system for the evaluation of renal function in post - Transplant Patients," *J. Eval. Clin. Pract.*, vol. 26, no. 4, pp. 1224–1234, 2020.
- [37] A. Hajian and P. Styles, "Application of Neuro-Fuzzy Systems in Geophysics," in *Application of Soft Computing and Intelligent Methods in Geophysics*, Springer, 2018, pp. 417–484.
- [38] G. Robins, T. Snijders, P. Wang, M. Handcock, and P. Pattison, "Recent developments in exponential random graph (p\*) models for social networks," *Soc. Networks*, vol. 29, no. 2, pp. 192–215, 2007.

# A Privacy-Centered Protocol for Enhancing Security and Authentication of Academic Certificates

Omar S. Saleh<sup>1</sup>, Osman Ghazali<sup>2\*</sup>, Norbik Bashah Idris<sup>3</sup>

Studies-Planning and Follow-up Directorate, Ministry of Higher Education and Scientific Research, Baghdad, Iraq<sup>1</sup>

School of Computing, Universiti Utara Malaysia, Kedah, Malaysia<sup>1,2</sup>

Kulliyah of Information and Communication Technology, International Islamic University Malaysia, Kuala Lumpur, Malaysia<sup>3</sup>

**Abstract**—Academic certificate authentication is crucial in safeguarding the rights and opportunities of individuals who have earned academic credentials. This authentication helps prevent fraud and forgery, ensuring that only those who have genuinely earned certificates can use them for education and career opportunities. With the increased use of online education and digital credentials in the digital age, the importance of academic certificate authentication has significantly grown. However, traditional techniques for authentication, such as QR code, barcode, and watermarking, have limitations regarding security and privacy. Therefore, proposing a privacy-centred protocol to enhance the security and authentication of academic certificates is vital to improve the trust and credibility of digital academic certificates, ensuring that individuals' rights and opportunities are protected. In this context, we adopted the Challenge Handshake Authentication (CHA) protocol to propose the Certificate Verification Privacy Control Protocol (CVPC). We implemented it using Python and Flask with a Postgres database and an MVT structure for the application. The results of the implementation demonstrate that the proposed protocol effectively preserves privacy during the academic certificate issuance and verification process. Additionally, we developed a proof of concept to evaluate the proposed protocol, demonstrating its functionality and performance. The PoC provided insights into the strengths and weaknesses of the proposed protocol and highlighted its potential to prevent forgery and unauthorised access to academic certificates. Overall, the proposed protocol has the potential to significantly enhance the security and authenticity of academic certificates, improving the overall trust and credibility of the academic credentialing system.

**Keywords**—Academic certificates; privacy-centered protocol; privacy preservation; challenge handshake authentication protocol

## I. INTRODUCTION

Academic certificates are important documents that verify an individual's educational achievements and qualifications, serving as proof of their knowledge, skills, and competencies in a particular field of study. Employers, educational institutions, and other organisations often use academic certificates to evaluate an individual's qualifications for a job or further education opportunities. A study by the National Center for Education Statistics (NCES) found that in 2020, about 42% of the US population aged 25 and over had at least a bachelor's degree, which is an increase from 29% in 2000. This increase in the number of college graduates highlights the importance of

academic certificates, as employers and institutions require higher levels of education and skills [1], [2], [3].

Moreover, academic certificates play a crucial role in career advancement. A study by the Georgetown University Center on Education and the Workforce revealed that a worker with a bachelor's degree earns about \$1 million more in median lifetime earnings than a worker with only a high school diploma. In conclusion, academic certificates are important documents used to verify an individual's education and qualifications. They are becoming increasingly important as the number of college graduates increases, and employers and institutions require higher levels of education and skills, playing a vital role in career advancement. Academic certificate forgery is a severe issue affecting educational institutions, employers, and individuals [26], [31].

A study by the International Association for Educational Assessment (IAEA) suggests that academic certificate forgery is a growing problem worldwide, with an estimated 5-10% of all certificates being fraudulent. The rate of diploma fraud was 0.1% among the 3.5 million students whose records were checked in the United States, according to a study by the National Student Clearinghouse Research Center. The most common form of fraud was the use of a false high school diploma to gain admission to college [22]. In India, a study by the Centre for Media Studies found that up to 40% of engineering graduates in India may have fake degrees. The study revealed that many students could obtain fake degrees from unaccredited institutions or by paying bribes to officials. In China, a study by the China Academic Degrees & Graduate Education Development Center found that the rate of academic certificate forgery was about 3.15% among the population of college graduates, with most of the forgeries in the fields of engineering and medicine. While the methodologies of these studies may differ, they all suggest that academic certificate forgery is a widespread problem with severe consequences for educational institutions, employers, and society as a whole [1], [2], [3].

Therefore, it is essential for institutions and employers to have proper measures in place to detect and prevent certificate forgery to ensure that only qualified individuals are awarded and recognised for their education. Academic certificate verification is a critical process that helps mitigate the problem of academic certificate forgery by checking the authenticity of an academic certificate to ensure that a legitimate institution

\*Corresponding Author.

issued it and that the individual who holds the certificate completed the coursework and earned the degree.

A graduation certificate, also known as a diploma or degree certificate, is a document awarded to a student upon successfully completing a course of study or program, including the student's name, the name of the institution, and the degree or diploma earned. It typically includes the student's grades or other academic information in some cases. It serves as official proof of the student's educational achievements and is often required for further education or employment. Some universities, colleges, and vocational schools also provide an official transcript with the graduation certificate.

However, ensuring privacy in academic certificate issuance and verification systems is a critical concern [32]. The proposed Privacy-Centered Protocol for Enhancing the Security and Authentication of Academic Certificates is based on the Challenge Handshake Authentication (CHA) protocol and offers several key contributions that could revolutionise the academic certificate issuance and verification process. Firstly, the protocol prioritises the privacy of certificate holders by ensuring that personal information is not disclosed during the verification process. Secondly, it employs the CHA protocol, which uses a challenge-response mechanism to enhance the security of academic certificates. Additionally, the protocol is tamper-proof and uses digital signatures to ensure the authenticity of certificates. It is also easily accessible to all stakeholders and compatible with commonly used devices, making it a practical solution for academic institutions. Finally, the protocol is designed to be compatible with emerging technologies such as blockchain [8], ensuring its longevity and continued relevance. Overall, the proposed protocol focuses on privacy, security, tamper-proof nature, accessibility, and compatibility, making it a promising solution for enhancing the security and authentication of academic certificates. The next subsections will discuss the importance of graduation certificates and the related security and privacy requirements.

#### A. Importance of a Graduation Certification Verification (GCV) System

A Graduation Certification Verification (GCV) system is of paramount importance in the modern job market and academic landscape. The system provides a reliable and secure means of verifying the authenticity of academic certificates, which is essential for employers, academic institutions, and government agencies. With the increasing prevalence of certificate fraud, a GCV system is critical for maintaining the integrity of the certification process and preventing fraud. By ensuring that only authentic certificates are accepted, a GCV system can help improve the job market's quality, reduce the risk of hiring unqualified candidates, and maintain trust in the certification process. Moreover, a GCV system can help to simplify and streamline the certification process, saving time and reducing costs for employers and academic institutions. Overall, a GCV system is essential for ensuring the accuracy and integrity of academic certificates and maintaining trust in the certification process.

#### B. Security and Privacy Requirements of a Graduation Certification Verification (GCV) System

A Graduation Certification Verification (GCV) system requires high security and privacy to ensure the authenticity of academic certificates. The system must protect the privacy and security of certificate holders' personal information and the confidentiality of the certificate itself. The following are some of the security and privacy requirements of a GCV system:

- **Authentication and Authorisation:** The GCV system must implement strong authentication and authorisation protocols to ensure that only authorised stakeholders can access and modify data [27],[28].
- **Data Encryption:** The system should use encryption technology to protect data in transit and at rest, ensuring that data is not accessible to unauthorised parties [26],[27],[28].
- **Secure Storage:** The system must use secure storage mechanisms to protect the confidentiality and integrity of the data, preventing unauthorised access and data loss.
- **Privacy Protection:** The GCV system must ensure that the certificate holder's personal information is protected, including their identity and other sensitive information[26],[27],[28].
- **Data integrity:** The system must ensure that the data stored in the system is authentic and cannot be tampered with. This can be achieved by using techniques such as digital signature, hash functions, and encryption [5].
- **Non-repudiation:** The system must provide a mechanism to ensure that the certificate holder cannot deny their ownership of the certificate. This can be achieved by using digital signature and public key infrastructure [6].
- **Access control:** The system must have a mechanism to control who can access the certificates and what they can do with them. This can be achieved by using role-based access control, access control lists, and permission-based systems [7],[28].

Ensuring security involves safeguarding user and stakeholder privacy and preventing unauthorised access, use, modification, or destruction of data to maintain information confidentiality. This defines the system's ability to provide protection [29]. Preserving privacy is widely considered to rely on access control technology, which is regarded as the most vital aspect [30]. By implementing these security and privacy requirements, a GCV system can protect the integrity of the certification process, prevent fraud, and maintain the trust of stakeholders. Implementing these requirements can also help ensure compliance with data protection regulations, such as GDPR, HIPAA, and CCPA.

## II. LITERATURE REVIEW

The issuance and verification of academic certificates are critical processes in the academic world. The need for accurate, reliable, and secure verification of academic certificates has become increasingly important with the rise of online education and the prevalence of certificate fraud. This has led to the development of various techniques and technologies for academic certificate issuance and verification, which range from traditional methods to more modern and sophisticated approaches. In this literature review, we will explore the different techniques that have been proposed for academic certificate issuance and verification, including paper-based methods, electronic certificates, and digital signatures. We will examine the advantages and limitations of each approach, as well as their effectiveness in addressing the challenges of certificate fraud, privacy, and security. By understanding the various techniques and technologies used for academic certificate issuance and verification, we can gain insight into the future of this critical area of academic administration.

Traditional techniques for academic certificate issuance and verification have been used for decades, typically involving paper-based certificates and manual verification processes. While these methods have proven effective in many cases, they have limitations that have led to the development of more advanced technologies. In this literature review, we will explore the advantages and disadvantages of traditional techniques for academic certificate issuance and verification, including the use of paper-based certificates, manual verification processes, and the challenges of fraud prevention, privacy, and security.

Paper-based certificates have been the most widely used method for academic certificate issuance, and their validity has been verified by manual methods. However, this process is time-consuming, resource-intensive, and prone to error. Manual verification of certificates can be time-consuming and inefficient, and it may also require significant resources. Moreover, paper-based certificates are vulnerable to fraud, and they can be easily replicated or falsified. Despite the limitations of paper-based certificates, they are still in use, particularly in developing countries, where the lack of digital infrastructure and resources makes it difficult to implement more advanced technologies. In these situations, the use of paper-based certificates remains the only viable option, and efforts are being made to improve the security and validity of paper-based certificates.

QR codes are two-dimensional barcodes that can store large amounts of data in a small space, making them a popular choice for academic certificate issuance and verification. The use of QR codes can help to reduce the risk of fraud, simplify the verification process, and increase the efficiency of certificate issuance. QR has been used to store information such as the student's name, degree name, graduation year, and name of the University, making it easy for employers and institutions to verify the authenticity of the certificate[9],[10].

In [33], the research proposes a system for issuing degree certificates that includes a digital signature and a QR code tag. The QR code tag contains the graduate student's data, such as name, GPA, CGPA, and institution alias. The Higher

Education Certificate Authentication System (HECAS) generates the digitally signed QR code, which is sent to the central HECAS server for verification. A smartphone application is required to authenticate the certificate by scanning the QR code. The proposed system aims to provide a secure and efficient way of issuing and verifying degree certificates.

In [34], the research proposes a system for real-time student identity card authentication using a QR code and a smartphone scanner. The system generates a unique QR code containing a student's matriculation number and other details, which is embedded in the identity card. A software application pre-installed in the smartphone scanner functions as a QR scanner, allowing for quick and efficient authentication. The proposed system aims to enhance the quality of authentication and overcome the problem of location and connectivity issues. The research shows that the smartphone scanner is an effective and faster means of authentication compared to other traditional means. The system offers a promising solution to the lack of innovation in information technology, particularly in developing countries like Nigeria.

In [35], the paper proposes a barcode-based academic certificate authentication system that uses cloud-based services to enhance security and accessibility. The system generates a unique barcode for each certificate, which can be scanned and verified using a mobile application. The authors suggest that the system could help reduce fraud and improve the verification process for academic certificates.

In [36], the paper presents a QR code-based certificate authentication and verification system for higher education. Barcodes are used to improve security and accessibility. The system generates a unique QR code for each certificate, which can be scanned and verified using a mobile application. The authors note that the system could help prevent the production and distribution of fraudulent certificates.

In [37], the paper proposes a barcode-based certificate verification system for distance education. Barcodes are used to improve the security and efficiency of certificate verification. The system generates a unique barcode for each certificate, which can be scanned and verified using a mobile application. The authors suggest that the system could help reduce the time and cost associated with traditional certificate verification methods.

Using QR codes for academic certificate issuance and verification can pose some risks in terms of privacy and security. While QR codes offer a convenient and efficient way to verify certificates, they can also contain sensitive information that could be at risk of data breaches, hacking or misuse. For example, if the QR code contains personal data such as a student's academic history, it could be used for discriminatory purposes if it falls into the wrong hands. Moreover, QR codes can be easily replicated, potentially leading to fraudulent certificates being produced. To address these issues, it is important to take appropriate security measures such as encryption, access controls, and data privacy policies. Institutions or organisations may also need to use other technologies and methods to supplement QR codes, depending on their specific requirements [11], [12], [13], [14].

In [38], the researchers proposed a verification system based on watermarking that uses a combination of visible and invisible watermarks to authenticate digital certificates. The system uses an encryption scheme based on a secret key to ensure the security of the embedded watermark and employs a unique identifier to prevent the certificate from being duplicated. The system was tested on a sample set of certificates and demonstrated high accuracy in verification.

In [39], the researchers proposed a certificate verification system that uses a combination of QR codes and watermarks. The system embeds a unique watermark into each certificate that can be used to verify its authenticity. The system also includes a QR code that can be scanned to access additional information about the certificate holder. The system was tested on a sample set of certificates and showed high accuracy in verification.

While watermarking can offer a secure method for academic certificate issuance and verification, there are still potential drawbacks in terms of privacy and security. One of the main concerns is that someone with the right tools and knowledge can remove or alter watermarks, leading to fraudulent certificates being produced. Moreover, watermarking could lead to the possibility of certificate forgery, as attackers may be able to replicate the watermark and create counterfeit certificates. Embedding watermarks may also raise concerns regarding data privacy since the watermarks may contain personal information that could be accessed or misused. Therefore, it is important to complement watermarking techniques with additional security measures, such as encryption, access controls, and policies to protect the privacy of personal data.

In [40], the researchers proposed an RFID-based certificate verification system that uses a combination of hardware and software components. Each certificate is equipped with an RFID tag that contains a unique identifier and other relevant information. When the certificate is presented for verification, an RFID reader scans the tag and sends the data to a central server, which validates the information and returns a response indicating the certificate's authenticity. The system was tested on a sample set of certificates and showed high accuracy in verification.

In [41], the researchers proposed an RFID-based verification system that uses a unique identifier and a cryptographic key to authenticate digital certificates. The system employs an RFID reader to scan the certificate and transmit the data to a central server, which uses the cryptographic key to verify the authenticity of the certificate. The system was tested on a sample set of certificates and showed high accuracy in verification. The researchers also noted that the use of RFID technology can help prevent certificate fraud since the RFID tag is difficult to replicate or alter without the proper tools and knowledge.

However, there are potential drawbacks to using RFID technology for academic certificate issuance and verification. One concern is that the use of RFID technology could result in unauthorised access to personal data since the RFID tag contains sensitive information. Attackers may use unauthorised RFID readers to intercept the data or copy the RFID tag's

content. Furthermore, the use of RFID technology may raise concerns about data protection and privacy since RFID tags are capable of tracking individuals and monitoring their movements. Additionally, the cost of implementing RFID technology may be higher than other methods, which could be a barrier to adoption for some institutions. Therefore, it is essential to take appropriate measures to protect the privacy of personal data and prevent unauthorised access or interception of the RFID data.

However, watermarking relies on the visibility of the watermark, which can be difficult to detect in low-resolution images and can be removed through image manipulation [15], [16]. Table I shows the comparison between QR-Code, Barcode, watermarking and RFID.

TABLE I. COMPARISON BETWEEN QR-CODE, BARCODE, WATERMARKING AND RFID

Technique	Advantages	Disadvantages	References
QR Code	Can be read quickly and easily using a smartphone camera; Can store a large amount of data in a small. Can be easily integrated into existing systems.	Can be easily replicated or forged. Can be easily damaged or obscured, making it difficult to read.	[17], [18], [19]
Barcode	Can be read quickly and easily using a barcode scanner. Can store a limited amount of data in a small space. Can be easily integrated into existing systems.	Can be easily replicated or forged. Can be easily damaged or obscured, making it difficult to read.	[17], [18], [19]
Watermarking	Can be used to embed hidden information in the certificate that can be used to verify authenticity. Can be difficult to replicate or forge.	Can be easily damaged or obscured, making it difficult to read. Can be computationally expensive to create and verify.	[17], [18], [19]
RFID	Can be read quickly and easily using an RFID reader. Can store a large amount of data in a small space. Can be used for contactless authentication.	Can be easily replicated or forged. Can be easily damaged or obscured, making it difficult to read. Can be expensive to implement and maintain.	[17], [18], [19]

### III. THE PROPOSED DESIGN OF CENTRALIZED CERTIFICATE VERIFICATION PRIVACY CONTROL PROTOCOL (CVPC PROTOCOL)

The proposed design of the Centralized Certificate Verification Privacy Control (CVPC) protocol aims to address the shortcomings of traditional techniques such as QR codes, barcodes, watermarking and RFID for academic certificate issuance and verification. The CVPC protocol utilises a centralised server for certificate issuance and verification, which is responsible for maintaining the integrity and authenticity of the certificates. This centralised server is

responsible for generating and issuing digital certificates, as well as verifying the authenticity of the certificates when requested. The CVPC protocol utilises a combination of advanced cryptographic techniques, such as digital signatures and hash functions, to ensure the integrity and authenticity of the certificates. The digital certificates are issued with a unique digital signature, which is generated by the centralised server using the private key of the issuing institution. The digital signature ensures that the certificate has not been tampered with or modified in any way. The CVPC protocol also utilises a unique identification number, which is embedded in the digital certificate and is used to verify the authenticity of the certificate. This identification number is generated by the centralised server and is based on the student's personal information, such as their name, date of birth, and the institution they graduated from. The CVPC protocol also includes a privacy-preserving mechanism, which allows the student to control who has access to their personal information and the digital certificate. The student is provided with a private key, which is used to encrypt the personal information and digital certificate. The private key is stored on the student's device and is only accessible by the student.

With this proposed solution, three main objectives were sufficed. Primarily:

- The privacy aspect. The student can lock and unlock their certificates.
- Only authorised entities can exist in the system with sufficient authorisation and access control.
- Timely verification of certificates by third parties.

From the previous details, privacy was a major drawback in most of the existing solutions when it came to certificate verification. The process itself is quite troublesome, especially for verifiers. The first components towards sufficing the objectives of this research start with developing a privacy-first solution that allows the user to control how the public views the certificate in a centralised environment. Fig. 1 shows the design for the proposed centralised certificate verification privacy control protocol.

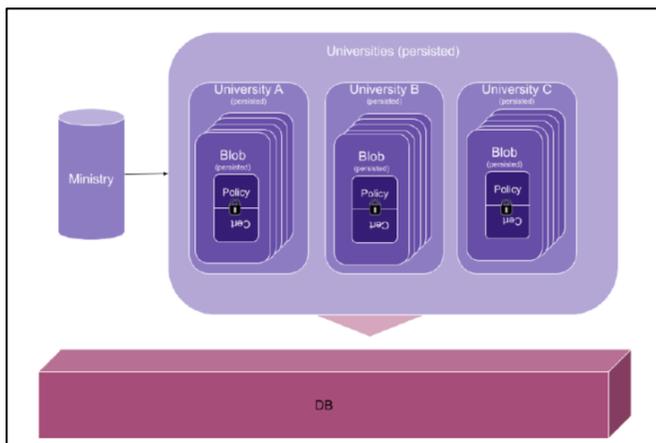


Fig. 1. The proposed centralised certificate verification privacy control protocol.

The number of ministries in the world is finite  $n(\text{Ministry}) = x, \{x = \text{constant} \mid x > 0\}$  and usually there is one per country  $\text{Ministry}_x \subset \text{Country}_x$  which is usually responsible for the quality of education in their respective countries. The ministry of education can go under different names like the case of Mexico it is called Secretariat of Public Education [16]. Nevertheless, this should not have any impact on the design since the ministry is an entity that controls and generates universities  $\{\text{University}_1, \dots, \text{University}_n\} \in \text{Ministry}_x \mid n \geq 1$ . The set of ministries based on the proposition above is finite hence they are hard coded into the system (by their official admin emails); Each Ministry is able to only create universities in the system. Each university is created by one and only one ministry. For the cases where there are different branches of a single university in different countries, this would neither be impacted nor will it cause the system to behave wrongly. For example, Birmingham University, originally situated in the UK and precisely England, would be created by the Ministry of Education in the UK. Birmingham University also has a branch in Malaysia. The Malaysian Ministry of Education would create the latter. Each university issues its own certificates. However, in the proposed solution, as a certificate is generated a privacy policy is also generated and attached to that generated certificate as shown in the design Fig. 1. By default, the certificate is locked, which means any third party attempting to view the certificate will not be able to see its details. When the university creates a certificate, another task is triggered in the background, generating a student for that certificate. The way this is achieved is by using the registered email of the student in that university. This is to suffice for authentication. The student would get an activation token that allows them access to the certificate and policy. The student after activating their account would be able to control how each piece of information is displayed to third parties. The mechanism used to achieve the above is an adaptation from the Challenge Handshake Authentication Protocol (CHAP) which is shown in Fig. 2. CHAP suffices when a link is between a server and a client [20],[23],[24].

- The server sends a challenger message to the client.
- The client responds.
- The server checks if the response matches the expected value, then the authentication is acknowledged, and the connection happens otherwise it is terminated.

The Challenge Handshake Authentication protocol was adopted in the proposition of the certificate verification privacy control protocol (CVPC). CHAP operates by first establishing a connection between the user and the network resource. The network resource then sends a challenge message to the user, typically a random string of characters or a nonce, which the user must use to generate a response message. The user generates the response by running a one-way hash function on the challenge message using a shared secret key known only to the user and the network resource. The resulting hash value is then sent back to the network resource, which compares it to its own calculation of the expected response. If the two values match, the user is authenticated and granted access to the network resource. Details are shown in the next section. The Challenge Handshake Authentication Protocol (CHAP)

proposed by [20],[21], which was discussed in the previous section, was adopted in that same manner in the CVPC protocol. It is used when a ministry adds a university, and a university issues a certificate and assigns a student. The adoption and the implementation of Challenge Handshake Authentication protocol (CHAP) in academic certificates verification use case is given in Fig. 3. It shows how the relationship between the university and the students.

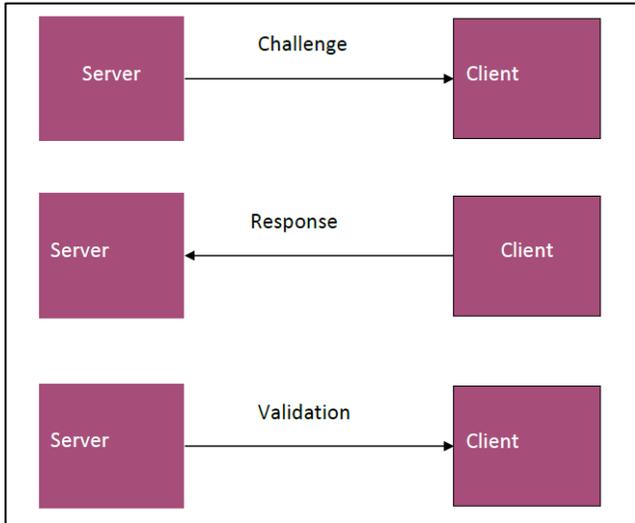


Fig. 2. Challenge Handshake Authentication Protocol (CHAP).

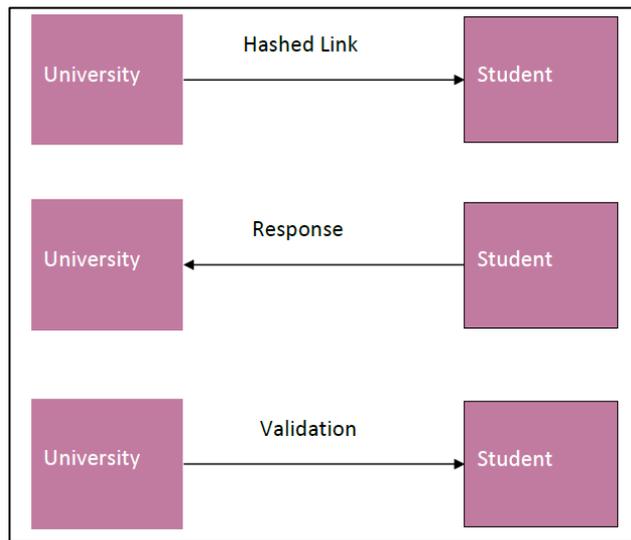


Fig. 3. Implementation of Challenge Handshake Authentication Protocol (CHAP) in academic certificates verification use case.

The proposed CHAP Certificate Verification Privacy Control Protocol used is based on the following steps:

Step 1: A request is sent to the student's email. The request is in the form of a signed link that is specific to that student, and the link contains a token. The token results from hashing of a user's identifier with the student's email and a timestamp. The timestamp is important to validate the lifespan of the link. The link is meant to be active for a pre-defined period of time

to minimise hanging certificates. Meaning certificates in the system without confirmed students.

Step 2: The students from their official emails access the link supplied. This step acts as a response to the request sent by the server. It is important to note that each link is unique to the student and can only be accessed from their official email, which requires authorisation and access control.

Step 3: The server validates the authenticity of the token sent by the student by decrypting it and retrieving the email address hashed in the token, using the same student identifier that was used to sign the token. If the token is valid, the student is directed to the certificate and policy and can unlock the certificate.

The proposed protocol uses HMAC and SHA-512 to sign the hashed link. HMAC stands for Keyed-Hashing for Message Authentication and is a widely used message authentication code based on cryptographic hash functions like MD5 and SHA-1. The student identifier acts as a namespace for the URL and is unique to each student. Therefore, the student identifier is usually the student's username or ID. This adds an extra layer of privacy since no two tokens can be decrypted using the same parameters. Since each email and student identifier is unique, each link is specific to a single student.

Based on what has been discussed in the previous section the following protocol is proposed (CVPC protocol).

Step 1: The ministry adds a university using CCVPC proposed,  $CCVPC(University) \in Ministry$ .

Step 2: The university adds students using CCVPC,  $University \rightarrow CCVPC(Student_n)$ ,  $n \in \text{list of students in university}$ .

Step 3: The university issues certificates with Privacy Policy,  $University \rightarrow PsxCs$ ,  $P=Policys$ ,  $C=Certificates$ ,  $s=\{Student_0, \dots, Student_n\}$ .

Step 4: Privacy Policy adds a layer of protection to the Certificates  $Ps(Cs)$ .

Step 5: All inbound traffic hits the Privacy Policy first.  $Inboundrequest \rightarrow Ps \rightarrow Cs$ .

Step 6: Student lock/Unlock their certificates sufficing privacy  $L(Cs)$ ,  $U(Cs)$ ,  $L=Lock$ ,  $U=Unlock$ .

Step 7: Students share ids of certificates with the third parties and based on the privacy policy they are able to see the information  $T(PsCs) \ni Student_n \Rightarrow U(Cs)$ ,  $T=Third\ party$ .

#### IV. THE PROPOSED VERIFICATION OF CERTIFICATES BY THIRD PARTIES

The third party can simply use a fixed id supplied by the student to access the certificate and validate the information.  $T(PsCs) \ni Student_n \rightarrow U(Cs)$ ,  $T=Third\ party$ ,  $Ps=Student\ Policy$ ,  $Cs= Student\ Certificate$ ,  $U=Unlocking$ .

The student can after being authorised into the system using the CCVPC proposed will be able to control each element of his/her certificate. Such that.

Studentn → L(Cei), ei ∈ Certificates | Student=Studentn where.

ei represents several elements like the GPA, transcript and other necessary information that is issued with the certificate. The CCVPC is a component of the proposed centralised certificate verification privacy protocol. This design ensures that the system remains free of unwanted parties. Access control is managed by authorised entities, each of which has their own control rights in the system. For example, universities issue certificates, but can only participate in the system through ministries. This control mechanism ensures proper authorisation, verification, and limits fraud. In addition, the CCVPC protocol ensures privacy by controlling who has access to the L(Cs) and U(Cs) of certificates - only the students themselves have this access. Thanks to the CCVPC protocol, no unauthorised entities can exist within the system.

## V. IMPLEMENTATION AND RESULTS

The proposed design was implemented in Python, a widely-used programming language that is particularly effective for large-scale web applications. The web framework Flask was used with Python, while the database was implemented with Postgres, an open-source database that natively supports JSON objects. The application was structured according to the MVT architecture, which stands for Models, Views, and Templates. Models describe the database, while Views implement the business logic, and Templates provide the front-end interface with HTML and CSS. In the following subsections, we will explore the different layers of the application in more detail.

- common
- models
- resources
- static
- templates
- app.py
- blacklist.py
- config.py
- db.py

### A. The Model Layer

The model folder contains all the entity models that were defined. These models represent the various actors involved in the use case, with the exception of the third party. Since third parties do not require an identity in the system, they can simply use the ID provided by the student to facilitate the verification process. In this process, several actors are involved, including the ministry, the university, students, and third parties. The ministry's responsibility is to add universities to the system, while the university takes charge of certificate issuance. Students have the ability to lock and unlock their certificates for privacy preservation. Finally, third parties are responsible for verifying the certificate. In the proposed protocol design, there are several actors involved, as depicted in Fig. 4. Additionally, this research includes the implementation of the Ministry class in Python, as demonstrated in Fig. 5.

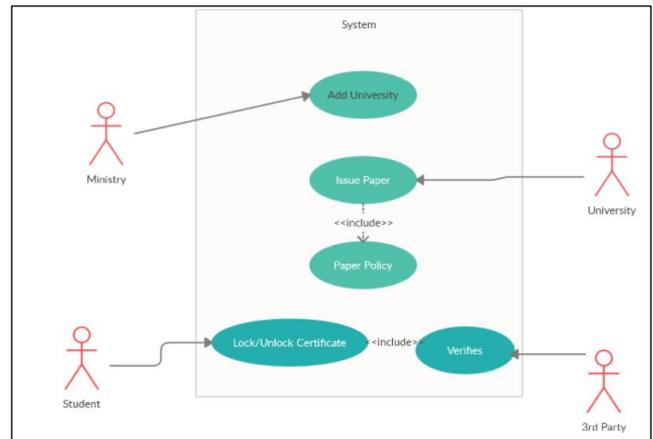


Fig. 4. Actors involved in the proposed protocol.

```

class MinistryModel(db.Model):
    __tablename__ = "ministries"
    __table_args__ = (UniqueConstraint("id"),)
    id = db.Column(db.Integer, primary_key=True)
    is_verified = db.Column(db.Boolean, default=False)
    name = db.Column(db.String(120), unique=True)
    country = db.Column(db.String(120), unique=True)
    password = db.Column(db.String(128))
    
```

Fig. 5. The implementation of the ministry class in python.

The same principle applies to both universities and students, with an additional class of permissions that facilitates control over the privacy of certificates. This permission class is linked to the student class through the use of student IDs.

### B. The View Layer

The logic that connects the front end to the model layer is implemented in this layer. Specifically, the ministry resource includes the following resources:

- Ministry(Resource).
- MinistryRegistration(Resource).
- MinistryList(Resource).

Each of the resources listed above serves a specific purpose. For example, the ministry resource exposes an API call that returns details about the ministry associated with a given email address.

- class Ministry(Resource):
- def get(self, email):
- email = email.lower().strip().
- Ministry = MinistryModel.find\_by\_email(email).
- if Ministry:
- return Ministry.json().
- return {"message": "Ministry not found"},

The Ministry Registration resource takes in the necessary information to create a new ministry. The creation of a ministry is pre-defined. The pre-defined entities are added using the post method exposed by the Ministry Registration resource above. The last Resource MinistryList(Resource) allows super admin to list existing ministries or it can also serve the general query the list of ministries registered in the system.

### C. Application Front End Layer

The student is able to control the permission using the endpoint /permissions/<int:student\_id>.

In this section, the proposed system's app screens illustrate various tasks, including adding a university to the system, the university's view for adding a certificate, the student's view for managing permissions, and the third-party's view of the application. Fig. 6 displays a screenshot of the university addition process, while Fig. 7 showcases a screenshot of the university adding a certificate. Furthermore, Fig. 8 demonstrates how a student can preserve their privacy. The student view for managing permissions is shown in Fig. 9, which displays a screenshot of a third-party verifying a certificate shared by the student.

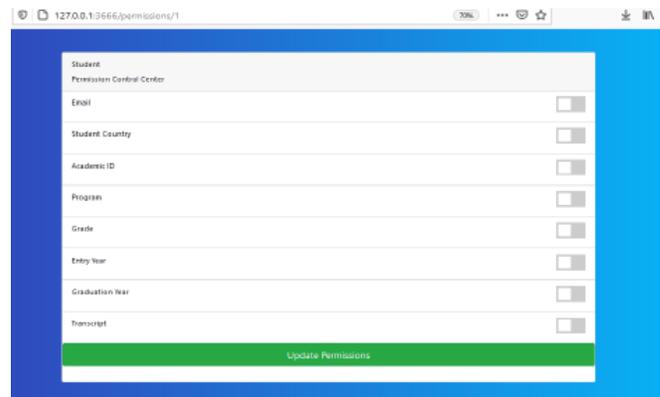


Fig. 8. Screenshot of how a student can preserve their privacy.

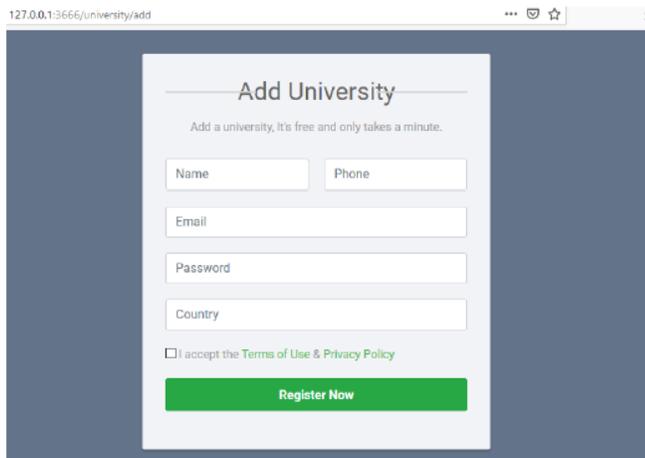


Fig. 6. Screenshot of adding a university.

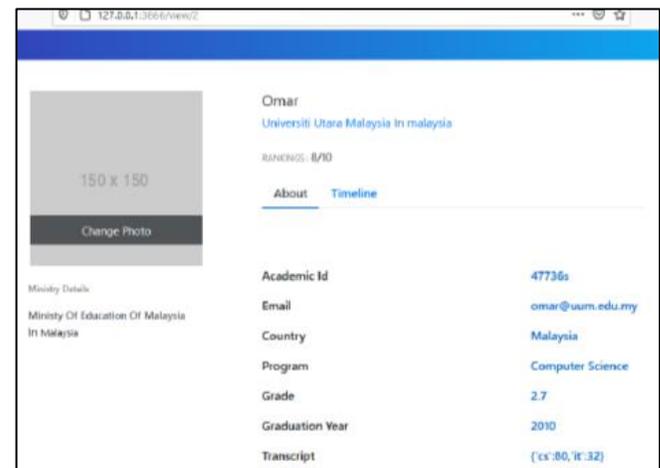


Fig. 9. Screenshot of third-party verifying the certificate shared by the student.

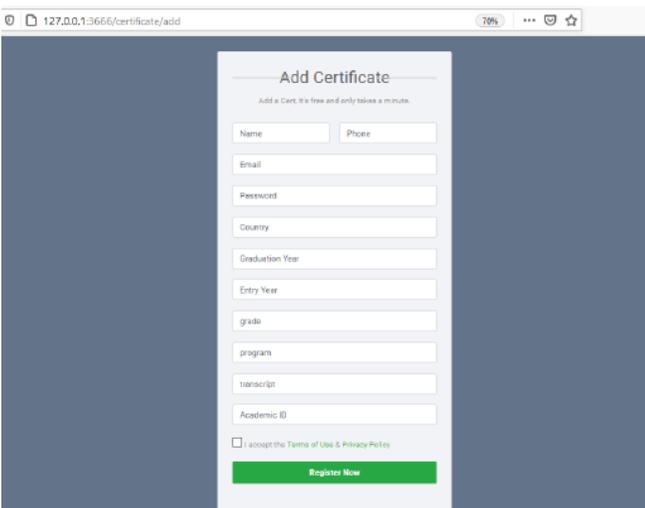


Fig. 7. Screenshot of adding a certificate by the university.

## VI. DISCUSSION

The Privacy-Centered Protocol for Enhancing the Security and Authentication of Academic Certificates based on the Challenge Handshake Authentication (CHA) protocol has the potential to significantly enhance the security and authenticity of academic certificates. The protocol's key contributions, including its focus on privacy, tamper-proof nature, and compatibility with emerging technologies, make it a promising solution for academic institutions. One of the most significant benefits of the proposed protocol is its emphasis on privacy. The protocol ensures that personal information is not disclosed during the verification process, providing an extra layer of protection to certificate holders. This privacy-centric approach is essential in today's digital age, where data breaches and identity theft have become significant concerns for individuals and organisations alike. Another significant advantage of the proposed protocol is its tamper-proof nature. The use of digital signatures ensures the authenticity and integrity of academic certificates, making it easy to detect fraudulent certificates. This tamper-proof nature can significantly enhance the overall trust in the academic certificate issuance and verification process, providing a more secure means of verifying academic credentials. The protocol's compatibility with emerging technologies such as blockchain is also noteworthy. The integration with blockchain technology can further enhance the

security and reliability of academic certificate verification. Blockchain technology provides a decentralised and tamper-proof way of storing and verifying data, making it an ideal solution for enhancing the security and authenticity of academic certificates. Despite the potential benefits, the adoption of the proposed protocol may face challenges. One of the challenges is the adoption and integration of the protocol into existing systems. It may require significant changes to the existing infrastructure and systems, which can be time-consuming and costly. Additionally, compatibility with emerging technologies such as blockchain requires a certain level of technical expertise, which may be a barrier to some organisations. In conclusion, the Privacy-Centered Protocol for Enhancing the Security and Authentication of Academic Certificates based on the CHA protocol is a promising solution for enhancing the security and authenticity of academic certificates. The protocol's emphasis on privacy, tamper-proof nature and compatibility with emerging technologies makes it a practical and robust solution for academic institutions. While there may be challenges to its adoption, the benefits of adopting the protocol far outweigh the potential challenges, providing a more secure and reliable means of verifying academic credentials.

#### VII. BENEFITS OF PRIVACY CENTRALISED VERIFICATION CONTROL PROTOCOL FOR ACADEMIC CERTIFICATES ISSUANCE AND VERIFICATION

Proposing a new privacy-centralised verification control protocol for academic certificates issuance and verification is important for several reasons [25], including:

- **Security:** Traditional methods of academic certificate authentication, such as QR code and barcode, are vulnerable to tampering and replication. A new privacy-centralised verification control protocol can enhance the security of the certificate verification process by incorporating advanced security techniques such as digital signature, encryption, and biometrics.
- **Privacy:** Centralised certificate verification systems can be a potential privacy breach, as they can expose the personal information of certificate holders to unauthorised access. The new privacy-centralised verification control protocol aims to preserve the privacy of the certificate holders by implementing a privacy-preserving protocol that protects the personal information of the certificate holders from unauthorised access.
- **Scalability:** The proposed protocol can handle a large number of requests and users, which is crucial in today's digital era where the use of digital credentials has increased.
- **Compliance:** The proposed protocol can ensure compliance with various privacy regulations and standards, such as GDPR, which is increasingly important as organisations have to comply with more stringent regulations to protect personal data.

In summary, proposing a new privacy-centralised verification control protocol for academic certificate issuance

and verification is crucial to ensure the security, privacy and scalability of the system, as well as to provide an efficient user experience and compliance with regulations.

#### VIII. EVALUATION OF THE PROPOSED PROTOCOL

Evaluating a proposed design of the Centralized Certificate Verification Privacy Control Protocol (CVPC Protocol) is crucial to ensure its effectiveness and efficiency in improving the security and privacy of academic certificate authentication. One way to evaluate the proposed protocol is by developing a proof of concept (PoC) and testing it with real-world scenarios and data. The PoC can be used to demonstrate the functionality and performance of the proposed protocol and provide insights into its strengths and weaknesses. The PoC development process can involve several steps, such as designing the system architecture, implementing the proposed security measures and privacy-preserving protocols, and testing the system with simulated or real-world data. The PoC can be evaluated based on various performance metrics, such as security, privacy, scalability, and usability. For example, the security of the proposed protocol can be evaluated by assessing its resistance to various security threats, such as tampering and replication. The proposed protocol's privacy can be evaluated by examining its compliance with various privacy regulations and standards, such as the General Data Protection Regulation (GDPR). Similarly, the protocol's scalability can be evaluated by assessing its capacity to accommodate a significant number of requests and users. Additionally, the usability of the proposed protocol can be evaluated by analysing its user interface and user experience.

#### IX. CONCLUSION

In conclusion, the proposed Design of Centralised Certificate Verification Privacy Control Protocol (CVPC Protocol) addresses the need for improved security and privacy in the realm of academic certificate authentication. The implementation of the proposed protocol involved the use of several technologies, including Python, Flask, and a Postgres database, as well as the utilisation of an MVT structure. Through the utilisation of these technologies and methodology, the proposed protocol has effectively demonstrated the preservation of privacy throughout the academic certificate issuance and verification process. A proof of concept was developed to further validate the functionality and performance of the protocol, which revealed its potential to prevent certificate forgery and unauthorised access. The CVPC Protocol proposed presents a promising solution for improving the security and privacy of academic certificate authentication. Future work involves building the protocol based on a blockchain platform.

#### ACKNOWLEDGMENT

This research was supported by Ministry of Higher Education (MoHE) of Malaysia through Fundamental Research Grant Scheme (FRGS/1/2018/ICT04/UUM/02/17).

#### REFERENCES

- [1] Rios, J. A., Ling, G., Pugh, R., Becker, D., & Bacall, A. (2020). Identifying critical 21st-century skills for workplace success: A content analysis of job advertisements. *Educational Researcher*, 49(2), 80-89.

- [2] Protopsaltis, S., & Baum, S. (2019). Does online education live up to its promise? A look at the evidence and implications for federal policy. Center for Educational Policy Evaluation, 1-50.
- [3] Abelho, M., Fernandes, S., Mesquita, D., Seabra, F., & Ferreira-Oliveira, A. T. (2020). Graduate employability and competence development in higher education—A systematic literature review using PRISMA. *Sustainability*, 12(15), 5900.
- [4] Eaton, S. E., & Carmichael, J. J. (2023). Fake degrees and credential fraud, contract cheating, and paper mills: Overview and historical perspectives. *Fake Degrees and Fraudulent Credentials in Higher Education*, 1-22.
- [5] M. Gariup and J. Piskorski, "The challenge of detecting false documents at the border: Exploring the performance of humans, machines and their interaction," *International Journal of Critical Infrastructure Protection*, vol. 24, pp. 100–110, 2019, doi: 10.1016/j.ijcip.2018.10.005.
- [6] J. K. Adjei *et al.*, "Document Authentication System Preventing and Detecting Fraud of Paper Documents," *ProQuest Dissertations and Theses*, vol. 5, no. 2, pp. 58–63, 2014, doi: <http://dx.doi.org/10.1108/AP-05-2012-0049>.
- [7] N. Massing and S. L. Schneider, "Degrees of competency: the relationship between educational qualifications and adult skills across countries," *Large Scale Assess Educ*, vol. 5, no. 1, Dec. 2017, doi: 10.1186/s40536-017-0041-y.
- [8] E. Tijan, S. Aksentijević, K. Ivanić, and M. Jardas, "Blockchain technology implementation in logistics," *Sustainability (Switzerland)*, vol. 11, no. 4. MDPI, Feb. 01, 2019, doi: 10.3390/su11041185.
- [9] G. Grolleau, T. Lakkhal, and N. Mzoughi, "An introduction to the Economics of Fake Degrees," *J Econ Issues*, vol. 42, no. 3, pp. 673–693, 2008, doi: 10.1080/00213624.2008.11507173.
- [10] Aini, Q., Rahardja, U., Tangkaw, M. R., Santoso, N. P. L., & Khoirunisa, A. (2020). Embedding a blockchain technology pattern into the QR code for an authentication certificate. *Jurnal Online Informatika*, 5(2), 239-244.
- [11] Suteja, B. R., Imbar, R. V., & Johan, M. C. (2020). e-Certificate system based on Portable Document Format and QR Code for Academic Activities. *International Journal of Computer Science Issues (IJCSI)*, 17(6), 87-91.
- [12] Mayowa, O. O., Adedayo, E. W., Olamide, O. O., Awokola, J. A. P., & Sodipo, Q. B. (2021). Design and Implementation of a Certificate Verification System using Quick Response (QR) Code. *LAUTECH JOURNAL OF COMPUTING AND INFORMATICS*, 2(1), 35-40.
- [13] Mayowa, O. O., Adedayo, E. W., Olamide, O. O., Awokola, J. A. P., & Sodipo, Q. B. (2021). Design and Implementation of a Certificate Verification System using Quick Response (QR) Code. *LAUTECH JOURNAL OF COMPUTING AND INFORMATICS*, 2(1), 35-40.
- [14] Chanda, D. (2019). Barcode Technology and its Application in Libraries. Akanbi, LM, Bashorun, MT, Salihu, UA, Babafemi, GO, Sulaiman, K., & Kolajo, SO (2019). Application of Barcode Technology in Landmark University Centre for Learning Resources, Omu-Aran Experience. *Library Philosophy and Practice (e-Journal)*. Retrieved from <http://digitalcommons.unl.edu>.
- [15] Ray, A., & Roy, S. (2020). Recent trends in image watermarking techniques for copyright protection: a survey. *International Journal of Multimedia Information Retrieval*, 9(4), 249-270.
- [16] VELIČKOVIĆ, Z., VELIČKOVIĆ, S., & MILIVOJEVIĆ, Z. (2021). „Application of Watermark in the Form of QR Code in COVID Certificate Validation “. *Journal of Mechatronics, Automation and Identification Technology JMAIT*, 6(2), 1-5.
- [17] Agrahari, A. K., & Varma, S. (2021). A provably secure RFID authentication protocol based on ECQV for the medical internet of things. *Peer-to-Peer Networking and Applications*, 14(3), 1277-1289.
- [18] Calderoni, L., & Maio, D. (2020, September). Lightweight Security Settings in RFID Technology for Smart Agri-Food Certification. In 2020 IEEE International Conference on Smart Computing (SMARTCOMP) (pp. 226-231). IEEE.
- [19] 13 Khan, R. A., & Lone, S. A. (2021). A comprehensive study of document security system, open issues and challenges. *Multimedia Tools and Applications*, 80(5), 7039-7061.
- [20] Sale, O. S., Ghazali, O., & Al Maatouk, Q. (2019). Graduation certificate verification model: a preliminary study. *International Journal of Advanced Computer Science and Applications*, 10(7).
- [21] Otuya, J. A. (2019). A Blockchain approach for detecting counterfeit academic certificates in Kenya (Doctoral dissertation, Strathmore University).
- [22] González-Gaudiano, E. J., Meira-Carrea, P. Á., & Gutiérrez-Bastida, J. M. (2020). Green Schools in Mexico and Spain: Trends and Critical Perspective. In *Green Schools Globally* (pp. 269-287). Springer, Cham.
- [23] Hussein, K. Q. (2019). Client Authentication By Selected Secure Password-Based On Image Using Challenge Handshake Authentication Protocol. *Iraqi Journal of Information Technology*. V, 9(3), 2018.
- [24] Ibrahim, A. S., & Hussein, K. Q. (2019). Client authentication by selected secure password-based on image using challenge handshake authentication protocol. *Iraqi Journal of Information Technology*.
- [25] Ayub Khan, A., Laghari, A. A., Shaikh, A. A., Bourouis, S., Mamlouk, A. M., & Alshazly, H. (2021). Educational Blockchain: A Secure Degree Attestation and Verification Traceability Architecture for Higher Education Commission. *Applied Sciences*, 11(22), 10917.
- [26] Saleh, O. S., Ghazali, O., & Rana, M. E. (2020). Blockchain based framework for educational certificates verification. *Journal of critical reviews*, 7(3), 79-84.
- [27] Saleh, O. S., Ghazali, O., & Idris, N. B. (2021, January). A New Decentralized Certification Verification Privacy Control Protocol. In 2021 3rd International Cyber Resilience Conference (CRC) (pp. 1-6). IEEE.
- [28] Din, I. U., Hassan, S., Almogren, A., Ayub, F., & Guizani, M. (2020). PUC: Packet update caching for energy efficient IoT-based information-centric networking. *Future Generation Computer Systems*, 111, 634-643.
- [29] Hashim, N. L., Yusof, N., Hussain, A., & Ibrahim, M. (2022). User experience dimensions for e-procurement: A systematic review. *Journal of Information and Communication Technology*, 21(4), 465-494. <https://doi.org/10.32890/jict2022.21.4.1>
- [30] Din, I. U., Guizani, M., Kim, B. S., Hassan, S., & Khan, M. K. (2018). Trust management techniques for the Internet of Things: A survey. *IEEE Access*, 7, 29763-29787.
- [31] Eaton, S. E., & Carmichael, J. J. (2023). Fake degrees and credential fraud, contract cheating, and paper mills: Overview and historical perspectives. *Fake Degrees and Fraudulent Credentials in Higher Education*, 1-22.
- [32] Pathak, S., Gupta, V., Malsa, N., Ghosh, A., & Shaw, R. N. (2022). Blockchain-Based Academic Certificate Verification System—A Review. *Advanced Computing and Intelligent Technologies: Proceedings of ICACIT 2022*, 527-539.
- [33] Ahmed, H. A., & Jang, J. W. (2017). Higher educational certificate authentication system using QR code tag. *Int. J. Appl. Eng. Res*, 12(20), 9728-9734.
- [34] Emmanuel, A. A., Adedoyin, A. E., Mukaila, O., & Roseline, O. O. (2020). Application of smartphone qrcode scanner as a means of authenticating student identity card. *International Journal of Engineering Research and Technology*, 13(1), 48-53.
- [35] Abbas, A. A. (2019). Cloud-based framework for issuing and verifying academic certificates. *Int. J. Adv. Trends Comput. Sci. Eng*, 8(6), 2743-2749.
- [36] Singhal, A., & Pavithr, R. S. (2015). Degree certificate authentication using QR code and smartphone. *International Journal of Computer Applications*, 120(16).
- [37] Goyal, S., Yadav, S., & Mathuria, M. (2016, September). Exploring concept of QR code and its benefits in digital education system. In 2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI) (pp. 1141-1147). IEEE.
- [38] Aini, Q., Rahardja, U., Tangkaw, M. R., Santoso, N. P. L., & Khoirunisa, A. (2020). Embedding a blockchain technology pattern into the QR code for an authentication certificate. *Jurnal Online Informatika*, 5(2), 239-244.
- [39] Wellem, T., Nataliani, Y., & Iriani, A. (2022). Academic Document Authentication using Elliptic Curve Digital Signature Algorithm and QR

- Code. JOIV: International Journal on Informatics Visualization, 6(3), 667-675.
- [40] Khalil, G., Doss, R., & Chowdhury, M. (2019). A comparison survey study on RFID based anti-counterfeiting systems. *Journal of Sensor and Actuator Networks*, 8(3), 37.
- [41] Kewale, P., Gardalwar, A., Vegad, P., Agrawal, R., Jaju, S., & Dabhekar, K. (2021, September). Design and implementation of RFID based e-document verification system. In *2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA)* (pp. 165-170). IEEE.

# A Systematic Literature Review on AI Algorithms and Techniques Adopted by e-Learning Platforms for Psychological and Emotional States

Lubna A. Alharbi

Department of Computer Science, University of Tabuk, Saudi Arabia, 71491

**Abstract**—Computers are becoming increasingly commonplace in educational settings. As a result of these advancements, a new field known as CEHL (Computing Environment for Human Learning) or e-learning has emerged, where students have access to a variety of services at their convenience. Using an e-learning platform facilitates more efficient, optimized, and successful education. They allow for personalized instruction and on-demand access to relevant, up-to-date material. These e-learning strategies significantly impact learners' emotional and psychological states, which in turn affect their abilities and motivations. Because of the learner's physical and temporal detachment from their tutor, encouraging learners can be challenging, leading to frustration, doubt, and ambivalence. The learner's drive to learn will be weakened, and their emotional and psychological state will be badly impacted as a result, both during and after the learning session. This research aimed to learn about the methods currently used by research facilities to analyze human emotions and mental states. The findings reveal that only e-learning has been used in education and other fundamental technologies, including machine learning, deep learning, signal processing, and mathematical approaches. A wide variety of e-learning-focused real-world applications make use of these methods. Each study subject is explained in depth, and the most frequently used methods are also examined. Finally, we provide a comprehensive analysis of the prior art, our contributions, their ramifications, and a discussion of our shortcomings and suggestions for future research.

**Keywords**—Psychological states; emotional states; e-learning; online platforms; solutions

## I. INTRODUCTION

Technology has become an integral part of our lives in the twenty-first century, prompting a reevaluation of fundamental beliefs on the part of professionals, educators, and students in order to re-design or re-engineer the educational and training infrastructure. In addition, these technology tools play a crucial role in enabling students and educators to reap its many benefits [1]. The education community is left with the difficult task of increasing the number of creative and original graduates while keeping costs down by applying cutting-edge technical and ecological methods [2]. This process has been a deliberate evolution from the traditional Gurukula methods to the present day, when the digital world has invaded the realm of education and revitalized the student body by providing them with a dynamic, interactive, and electronic environment on which to study. There have been many shifts in the previous few decades, and the meaning of "e-learning" had to adapt. Web

1.0 (the web of knowledge), Web 2.0 (the web of communication), Online 3.0 (the web of interaction), and now Web 4.0 (the web of integration) are all instances of the great progression seen in the world of the web.

## II. LITERATURE REVIEW

Due to the accessibility of new technologies and their capacity to generate and maintain stakeholders, e-learning has risen to the fore in today's ever-changing and dynamic online environment. The term "e-learning revolution" describes the widespread adoption of technological aids to education. To promote safe, cooperative, constructivist, and long-term knowledge exchange, the education sector and its allies hope to usher in an era of paperless learning made possible by technological advancements. As the pace of the technical education revolution quickens, it becomes more difficult for stakeholders to keep up with their commitments [2].

No concerns or pressures were placed on either the educational institutions or the students due to the annual 15.4% increase in e-learning worldwide [3]. However, this study was undertaken during COVID-19, and a lot has changed since then. As a consequence of worldwide limitation measures established to curb the spread of COVID-19 [5], more than 60 percent of students throughout the globe today get the majority or all of their education online, including lectures and a variety of assessments on numerous platforms.

### A. Psychology and Emotional States

Psychological and emotional health is crucially important in many aspects of daily life. A person's emotional state is the overall emotional tone of their personality (especially with regard to pleasure or dejection). While the nature of a state may change over time, the concept of a "psychological state" refers to a more stable mental situation.

1) *Types of emotional states*: Researchers in [15], have identified 27 distinct human emotions: awe, admiration, amusement, anger, calmness, confusion, disgust, empathy, excitement, fear, horror, enchantment, entrancement, joy, nostalgia, relief, romance, sadness, satisfaction, sexual desire, and surprise. Moreover, Fig. 1 shows how multimodal settings and sources might identify an individual's emotional state. In addition, Table I outlines the various sorts of feelings that are significant to the surroundings, which gives concrete instances of how each of the seven emotion types is evaluated and how they tend to act.

TABLE I. ENVIRONMENTALLY RELEVANT EMOTION TYPES [4]

Emotion Type	Appraisals	Action-Tendency
Self-condemning Emotions (Guilt, Shame, Embarrassment)	Own Norm violations	Correction (e.g., repair the environmental damage)
Other-condemning Emotions (Anger, Disgust, Contempt)	Others' norm violations	Punishment (e.g., punish those responsible for environmental destruction)
Self-praising Emotions (Pride)	Own positive norm deviations	Support oneself (e.g., in-group favoring pro-environmental behavioral intentions)
Other -praising Emotion (Elevation, Admiration, Awe, Being Moved, Gratitude, Love)	Others' positive norm deviations	Support the source (e.g., protect nature)
Other-suffering Emotions (Compassion, /Empathy, Emotional Contagion)	Other's suffering	Help those in need (e.g., victims of environmental destruction)
Threat-related Emotions (Fear, Anxiety, Hopelessness)	Anticipated negative consequences	Escape (e.g., fleeing from climate change)
Hedonistic Emotions (Joy, Pleasure, Amusement)		Reinforcement (e.g., enjoying car driving predicts car use)

2) *Types of psychological states:* The distinction between the two perspectives is: Extraversion: Libido that is directed outward is known as extroversion [7]. With an extrovert, the subject's interest in the object shifts in a positive direction. To be introverted is to direct one's libido inward toward the subject's own core. This fact conveys the nature of the subject-object relationship. The focus shifts away from the item and back to the subject. These four roles include: There are two senses: Synergy of Sensation and Intuition Separate evaluations: Feelings and ideas In particular, it is important to note how the perceiving functions can be effective for the diagnostic a subjective process to get in relation to patients and for creativity, while the thinking and feeling processes are related to rationality and can well serve to scientific ideas. Learning styles and personality types in medical school 5 Understanding the different personality types can help provide insight into how pupils' learning styles come into play. Whereas psychological types are highlighted in [6] and scientist pinpoints emotions in [15] also plays a significant role. Emotional and psychological visualization in e-learning environments shown in Fig. 2.

This psychodiagnostic test proposes highlighting four types of learning styles:

1) *Concrete experience:* In-the-moment reflection and problem-solving that prioritizes intuitive, emotional, and visceral processes over logical, scientific ones. The best learning environments for people with such strong relational and social abilities contain minimal organization, direct participation in real-world challenges, and a willingness to share personal information and perspectives.

2) *Reflective observation:* stressing observation and comprehension over application, with a tendency to grasp the

meaning of ideas and circumstances. Subjects who exhibit this form of learning are experts at identifying causation and deducing consequences. They exhibit composure, impartiality, and independent judgment by seeing the same problems from diverse perspectives.

3) *Abstract conceptualization:* skill in working with ideas and concepts in accordance with logical principles, using mostly rational thought rather than emotion in the learning process. There is a predisposition toward quantitative reasoning, planning, and design in these fields. Precision, discipline, analysis, and the organic arranging of conceptual systems are expressed as values by these topics.

4) *Active experimentation:* propensity to take action to influence reality (regarding situations or individuals). His philosophy emphasizes doing as opposed to thinking, which compels him to approach life with a strong dose of pragmatism, placing importance on how things work rather than their intrinsic worth or ultimate significance. People who have this skill can influence their environments to get what they want. These findings have the potential to illuminate medical treatment that takes into consideration patients' own awareness of their own preferences in learning styles, personality traits, and so on.

Fig. 1 and Fig. 2 depict the visualization of emotional recognitions in human from different sources.

**B. Learning Types**

Technology-based education can be referred to by a variety of names, including e-learning, m-learning, and d-learning [1]. e-Learning can replace conventional schooling or work in tandem with it (e-learning m learning). e-Learning goes under many names, including e-education, distant learning, and online education.

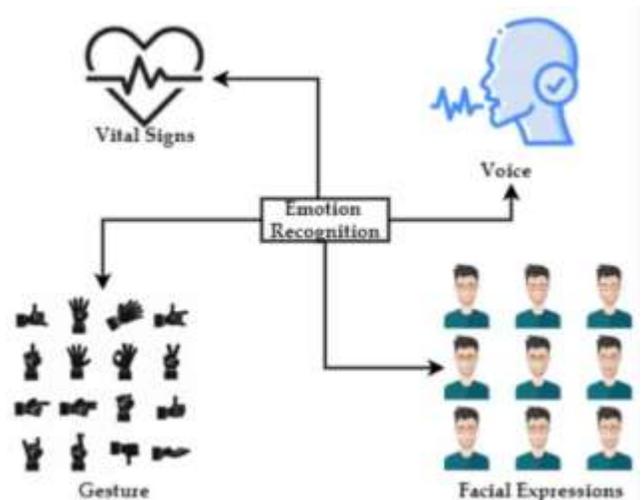


Fig. 1. Human emotional recognition of different sources [60].

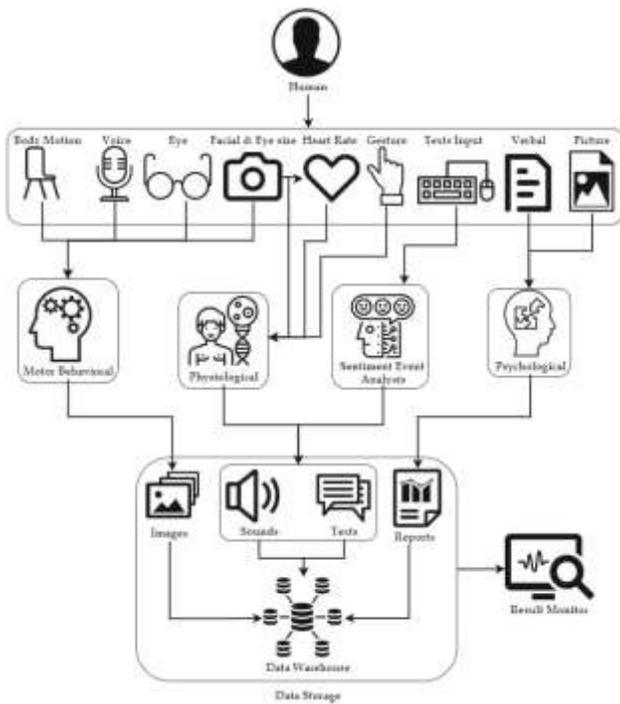


Fig. 2. Visualizations of emotional cartography for e-learning heuristic multimodal approaches [60].

The authors describe e-learning in [3, 8] as "the wide variety of applications and processes that leverage available electronic media and resources to offer vocational education and training." According to research [9], e-learning is "the use of multiple web-based, web-distributed, or web-capable technological instruments for education." Increasing numbers of individuals are becoming aware of the multiple advantages of e-learning [10], which include mobility, accessibility, and cheap cost. Considering these advantages, education may become a lifelong pursuit. According to [11], having limitless access to lectures assists students in retaining the essential information for formal education.

Higher education institutions are also adopting e-learning technology to expand the learning community and facilitate the flow of knowledge between students and teachers [12]. Due to its convenient scheduling, e-learning has the potential to attract more students who are otherwise unable to pursue higher education because of their other responsibilities at home or at work. In fact, this benefits not only the students, but also the teachers.

The founding of the National Center of e-learning and Distance Learning (NCEDL) in the Kingdom of Saudi Arabia in 2005 [13] involves at least nine institutions. This crucial role was created to enhance the overall e-learning experiences of students in schools by adopting and applying the finest e-learning techniques from across the globe [14]. According to the National Center for e-learning and Distance Learning, the NCEDL has participated in various e-learning system initiatives, including the Learning Portal, which gives students remote access to online learning resources and offers instructors training in the use of e-learning technologies.

To further encourage educational institutions to embrace e-learning, the center has created the Award for Excellence in e-learning, for which around 42 institutions are now competing. Since its inception in 2011, a large number of students and graduates have enrolled in courses at the Saudi Electronic University (SEU). Since then, King Abdelaziz University has developed several technological tools to enhance its e-learning system, including the Learning Management System (LMS), which provides access to over 16,000 e-books and other online academic materials for freshmen and juniors [14]. In addition, Tables II and III outline the kinds of e-learning systems, their major components, and their definitions.

TABLE II. E-LEARNING SYSTEM DEFINITION PRIOR RESEARCH [1]

Types of E-Learning System	Prior Research	Definition
Blended Learning	[16,17,18,20]	A mix of traditional and online classes.
Flipped Classroom	[17,20,21]	Focus on the individual learner by distributing preparatory readings and videos online.
ICT Supported Face-to-Face Learning	[22,23]	The integration of ICT with conventional teaching methods.
Synchronous Learning	[17,21,24]	A real-time interaction distance learning.
Asynchronous Learning	[17,21,24]	Non-real-time interaction distance learning.

TABLE III. E-LEARNING CRITERIA AND PRIOR RESEARCH [1]

Factors	Prior Research
Student Characteristics	[22,25-27]
Instructor Characteristics	[10,22,25-27]
Learning Environment	[10,22,25,27]
Instructional design	[10,22,25,27]
Support	[10,22,25-27]
Information Technology	[19,22,29,25,26,28]
Technology Knowledge	[10,22,25-27]
Course	[10,18,19,22,25,27]
Level of Collaboration	[10,19,26,27]
Knowledge Management	[10,19,26,27]

Researchers in [1] claim that e-learning, focusing on the online, is utilized in educational settings to teach and learn about a wide variety of electronic technologies (e.g., television, radio, CD-ROM, DVD, mobile phone, Internet, etc.). According to the definition, e-learning encompasses studying with web-based training facilities, such as digitally collaborative and technology-assisted distance learning offered by virtual universities and classrooms. It is possible to define e-learning innovation as any form of e-learning seen as a novel by its target audience, whether it be a new piece of technology or a new approach to teaching.

It's undeniable that e-learning has had and will continue to have a significant impact on educational progress around the globe. It also presents exciting new possibilities for developing countries eager to advance their educational infrastructure. In

addition, it facilitates the transition of the next generation of educators to the pedagogies of learning made possible by the digital age technologies. It's also been said that the internet and other modern technology are used to help education and training in ways that go well beyond the traditional classroom.

e-Learning, or education delivered by electronic means such as the Internet, CDs, DVDs, and mobile phones, arose in the 1980s as an alternative to traditional classroom instruction. Other benefits of online education have contributed to its rapid expansion in recent years. The following are some definitions of e-learning. e-Learning uses computer network technology to convey knowledge and instructions to humans, typically over the internet.

- The term "e-learning" refers to a wide range of uses and procedures, including using many online multimedia content delivery systems, including the World Wide Web, Internet video SD-ROMs, television, and radio. All of these resources are available to students to educate themselves.
- Web-based education, computer-based education, virtual classrooms, and digital collaboration are all examples of e-learning. Content can be distributed in a variety of ways, such as on the web, intranets, wide-area networks (WANs), CDs, DVDs, radio, television, satellite, and even cassette tapes.

“The experiential aspect of online education involves motivation, interest, experimentation, and repetition.

As indicated, the four significant e-learning perspectives illustrated in Fig. 3 are equally important in making electronic devices conceivable as tools for the delivery of educational institutions, and they are interconnected. The cognitive viewpoint analyses the function of the brain and its processes in learning from a logical standpoint. Smart learning systems and adaptive learning technology can be used to optimize learners' progress in an e-learning environment based on cognitive pedagogical models; similarly, virtual (simulated) worlds and other structured learning environments can facilitate students' comprehension of the subject matter.

Social media and other collaborative platforms can be used to facilitate conversation and learning through observation and imitation, and students can be coached through the use of the system in a short amount of time.

The emotional perspective considers the feelings of the learner and their environment. The researchers highlight several emotions as closely linked to integrating cognition, motivation, and behavior. These include pride, frustration, relief, resistance, fear, expectancy, hopelessness, anxiety, confidence, a complex, and jealousy.

Focus is placed on the skills and behavioral outcomes of the learning process from a behavioral viewpoint, emphasizing role-playing and practical application in real-world scenarios. Central to the contextual view is the learners' contacts with others, their discovery of the importance of collaboration, and the impact of their peers.

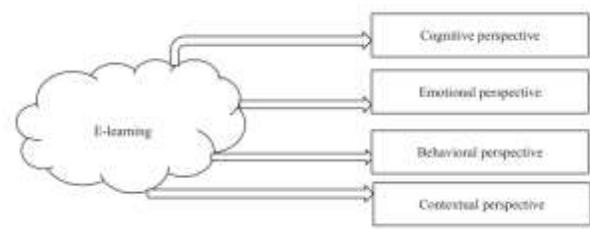


Fig. 3. Fundamental perspective of e-learning [1].

Table IV, illustrates the advantages and disadvantages of e-learning which focuses on the advantages of adopting it and as well as facing challenges by having it.

TABLE IV. ILLUSTRATES THE MOST COMMON ADVANTAGES AND DISADVANTAGES OF E-LEARNING [1]

Advantages	Disadvantages
<ul style="list-style-type: none"> <li>• Easy Access, individual instructions , different, flexibility, motivating and interesting, self-learning and self-improvement, feedback and evaluation, efficient and cost-effective strategy</li> </ul>	<ul style="list-style-type: none"> <li>• Required knowledge and skills, lack of equipment, isolation, missing social contact, negative attitude, technical defect, stressful and consumed more time, lack of co-curricular activities, lack of teacher training program</li> </ul>

### C. Challenges in e-Learning

Some difficulties arise from mediating technology [2] by putting forth consistent work overtime is the biggest obstacle to creating a learning company. Getting people interested in a new concept is straightforward, but consistently implementing it is far more challenging. When individuals are motivated and ready to learn, course content and organizational policies that are strategically aligned work together to use existing talent to accomplish corporate objectives. Administrators, instructors, and students of e-learning encounter a variety of hurdles. In the Web 0 era, students experienced a range of difficulties, including a fear of technology (7.24%) and bandwidth issues (3.0%); now, students suffer a lack of support from senior management (3.6 percent).

The current generation of e-learners has a new difficulty: getting businesses to recognize their degrees earned online. Web 0 implementation was problematic due to learners' epistemic beliefs and bandwidth availability. In seven pieces, educators from different generations of online education express their greatest worry about students' lack of willingness to learn. A designer's work is difficult due to the constant developments and upgrades of technology.

Dropout rates (11.97%) are a big obstacle for implementers. Still, dispersion in learner requirements (7.24%), synchronization of the most recent design and technology (5.07%), and unsuitable structural design (3.2%) have been significant issues for designers. 7.99 percent of firms described dealing with cultural opposition to be challenging. It has been observed that resistance to change is lessening as the Internet evolves. e-Learning stakeholders prioritized access to sophisticated technology and bandwidth for continued online course delivery (Diffusion of Innovation Theory) and the learning community's acceptance of online learning (Technology Acceptance Model).

The knowledge gap between the intended audience and the rest of the population filled by taking serious efforts. These efforts were made to make e-learning extremely interactive (Engagement theory). By keeping the goal of the learner interested and motivated (ARCS Theory). In today's world, students need to be engaged from the very beginning of a course if they are to remain motivated throughout its duration. This places a premium on the designer and implementer creating highly relevant, interactive, and individualized courses. One of the most important aspects of successful e-learning [2] is using the most appropriate and up-to-date technology for delivering the course.

This paper will adhere to the following structure. In Section III, we see an example of a research methodology with three major stages: review planning, review execution, and review reporting. Section IV presents the discussion. In Section V, the results of the chosen articles, study goals, standard processes, data formats, and performance approaches are discussed. Section VI discusses existing research, their contributions, managerial implications, and a conclusion that includes limitations and potential study pathways.

### III. RESEARCH METHODOLOGY

This Systematic Literature Review (SLR) methodology was based on the ideas presented in [30, 31]. Research is conducted in three distinct stages. As part of this preliminary preparation, we will discuss the steps of identifying research subjects, developing review procedures, and checking their accuracy. In the second, we discuss finding and choosing relevant research; in the third, we present the steps involved in writing and validating the SLR; and in the fourth, we discuss the process of information synthesis. Fig. 4 shows the progression of the three phases.

#### A. Plan Review

In this first stage of the research process, the relevant searching strategy is outlined alongside the key research questions and the establishment of review methods.

- *RQ #1*: What are the types of emotional and psychological states found and used in different types of learning?

The study's goal is to establish the utility of emotional and psychological states detected in the learning environment by organizations such as education sectors, development and training centers, and researchers for their models, frameworks, or applications. Nowadays, e-learning is being used, which has some emotional and psychological effects.

- *RQ #2*: Which type of algorithms and techniques admitted for the emotional and psychological states in e-learning platforms?

This study seeks to identify the approaches businesses, industries, and centers use in learning platforms such as online or e-learning and face-to-face learning.

- *RQ #3*: How emotional and psychological states observed/examined in e-learning platforms?

This study's subject is related to the algorithms, techniques, or models that are implemented in e-learning or face-to-face platforms, as well as identifying and evaluating the performance of these techniques in various e-learning platforms. This study aims to gain a comprehensive understanding of the procedures employed in Learning types and techniques. This review aims to look at models, frameworks, and applications that use e-learning and face-to-face approaches to address emotional and psychological difficulties.

1) *Review protocols*: The development and validation of the review protocol affirm the use of appropriate keywords to search for related articles and literature sources.

a) *Searching keywords*: To guarantee that the evaluation closely covers deep learning technologies for dental informatics, we attempted to focus our search to the most relevant specific keyword. As a result, we started with the terms and then proceeded to the next steps:

- i) Extracting the key terms from our study questions.
- ii) Using different terminology.
- iii) Adding keywords from relevant publications to our search terms.

As indicated in Table V, we used the primary alternatives and added the "OR operator" and "AND operator" to find the most immediately relevant works in the literature.

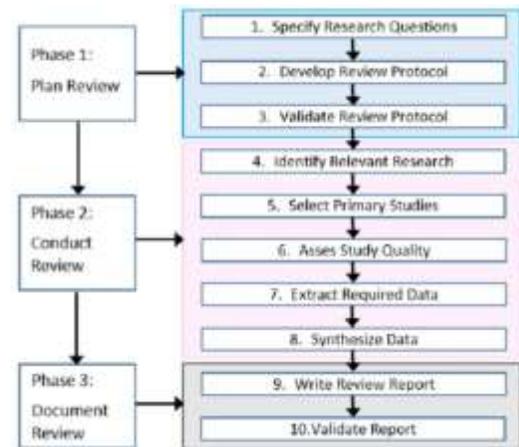


Fig. 4. SLR process.

TABLE V. INCLUSION AND EXCLUSION CRITERIA DESCRIPTION

ID	Keywords
1	("Psychological states" OR "psychological effect") AND ("Emotional states" OR "Emotions") AND ("Learning" OR "E-Learning")
2	("Psychological states" OR "psychological effect") AND ("Emotional states" OR "Emotions") AND ("Online Learning" OR "E-Learning")
3	("Psychological states" OR "psychological effect") AND ("Emotional states" OR "Emotions") AND ("Learning" OR "E-Learning") AND ("Tools" OR "Techniques")
4	("Psychological states" OR "Psychological effect") AND ("Emotional states" OR "Emotions") AND ("Learning" OR "Online Learning" OR "E-Learning") AND ("Tools" OR "Techniques")

b) *Literature resources:* The databases Web of Science, Scopus, ACM Digital Library, Springer, Science Direct, and IEEE Explorer were used to find relevant publications for primary review research. These databases, which include ISI, Scopus indexed papers, and publications from major conferences, provide the most comprehensive coverage of quality literature on our topic. The search phrase was developed by utilizing the extensive search possibilities provided by each of these databases. Our search included the years 2013 through 2022.

2) *Conduct review:* We used the research questions, keywords, and protocols as a reference to conduct the review in this step. This phase mostly deals with article inclusion and exclusion, as seen in (A) and (B) of Table VI.

TABLE VI. (A) INCLUSION CRITERIA DESCRIPTION AND (B) EXCLUSION CRITERIA DESCRIPTION

(A) Inclusion Criteria
The research was relevant to psychological and emotional states.
The research was directly related to the learning platforms.
The research was conducted using techniques and algorithms used by learning platforms.
The research is used in multiple domains.
The research was conducted for the analysis of algorithms and techniques performance in Learning Platforms.
For duplicate publications of the same study, the newest and most complete one was selected. This is recorded for only one study whose related work appeared two times.
(B) Exclusion Criteria
Studies unrelated to emotional and psychological states in dance, music, or any other field than education and health were excluded. Because traditional forecasts and visualizations are referred regarded as having "emotional and psychological effects," these results appeared in our search.

a) *Study selection:* Study selection is shown in its entirety in Fig. 5. There were a total of 1779 items found through the search. After sorting by title, keyword, and inclusion/exclusion criteria, we narrowed the list down to 150 articles. The criteria for inclusion and exclusion are listed in (A) and (B) of Table VI, respectively. Fifty-two papers were disqualified as a result of questionnaire-based predictions, and another 68 were disqualified because they dealt with other concepts, such as a theoretical model or a conceptual framework. Thirty items are crossed off the list after careful reading of the articles.

The selection criteria for relevant articles based on keywords are described in Table VI. Duplicate articles and those that do not address all of the research questions are omitted.

The quality checklist criteria for study evaluation are included in Table VII. The questions are primarily meant to assist in the selection of studies that are more relevant, thorough, and comprehensive in nature.

b) *Data extraction:* In order to obtain the data which are needed to address our research questions and contributions, we used the data-extraction methods highlighted in Table VIII.

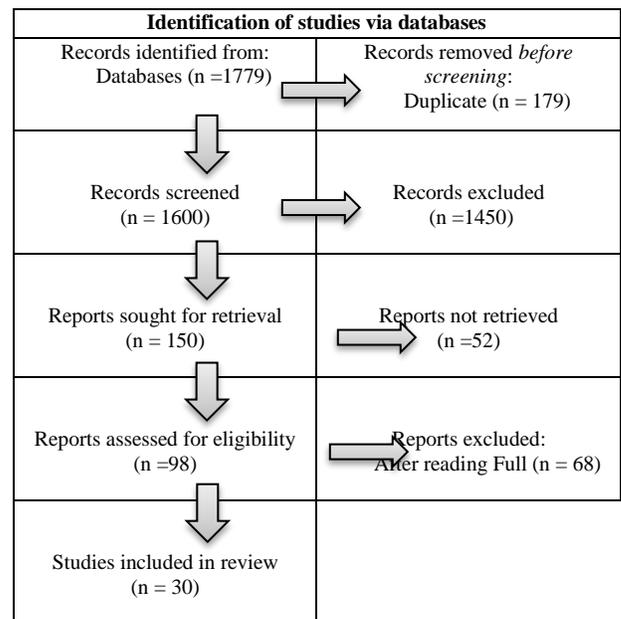


Fig. 5. Process of identifying relevant studies.

TABLE VII. QUALITY CHECKLIST

No.	Questions
1	Was there a strong focus on emotional and psychological states?
2	Was the study able to describe how emotional and psychological states is applied in learning?
3	Is there any algorithm or technique used to evaluate emotional and psychological states in education and medical learning that has been proposed?
4	Is the study concentrating on the basic learning approaches for Learning systems?
5	Is there any mention of core approaches used in the study?

TABLE VIII. DATA EXTRACTION

Study
Study Research Problem Contributions
RQ1: Psychological States
RQ2: Emotional States
RQ3: Techniques or algorithms used by learning platforms

c) *Information synthesis:* At this point, the retrieved data were pooled in order to respond to the research questions. For our research questions, we used the approach of narrative synthesis. Consequently, we used tables and graphs to describe our findings.

d) *Report review:* Four research questions were answered using information taken from primary studies. In describing the findings, strict adherence was made to the recommendations presented in [29,30].

#### IV. DISCUSSION

Emotions play an essential role in many facets of everyday life. We describe them as the predictable reactions we always have to unforeseen stimuli [32-37]. There is a short duration to these responses, which can be physical (muscle twitching,

trembling, etc.), behavioral (angry, fleeing, aggressive, immobile, etc.), physiological (sweating, redness, discomfort, pallor, accelerated pulse, palpitations, feeling ill, etc.), or psychological (positive or negative thoughts). Numerous studies have shown that emotions may alter the quality of learning if the learner's motivation is seen as a barrier to achievement [38-43]. Whether in a classroom setting, under test conditions, or in the comfort of one's own home, the process of learning is always accompanied with a complex and nuanced range of feelings [44-48]. Emotional influences on training in the workplace are the focus of this research. How emotions play a crucial role in learning, especially at a distance.

## V. RESULTS

In Table IX, 43 research met the criteria for inclusion. 15 research focused on emotional states utilized in e-learning platforms, and another 13 studies covering the various AI-based methods for assessing these states. Whereas 15 studies helped in addressing the question about the methods employed to assess mental health.

TABLE IX. RQ STUDIES

<i>RQ</i>	<i>Studies</i>
Emotional and psychological States	15
Emotional and psychological states evaluation techniques	13
Techniques or algorithms used by learning platforms	15

RQ #1: What are the types of emotional and psychological states found and used in different types of learning?

Effective feedback systems may aid in re-engaging and encouraging learners in these circumstances [32], which can eventually lead to enhanced learning. Therefore, in an e-learning situation, a successful system should be able to read the learners' emotions and evaluate their attention to deliver intelligent feedback that enhances the learners' learning experience. In e-learning circumstances, embodied conversational agents (ECAs) can give learners with effective and intelligent feedback. Sadly, creating these systems can be very difficult. Matching emotional states to facial expressions might be difficult when working with emotion recognition.

To overcome this, Paul Ekman attempted to map typical facial expressions for emotions like contempt, fear, fury, sadness, and surprise [34]. In addition, the effective identification of emotion by a computer in the late 1990s from IBM Watson was a significant milestone. Appropriate feedback systems may assist learners in regaining their footing and motivating them, eventually leading to enhanced learning. Embodied conversational agents (ECAs) [33] can provide effective and intelligent feedback in e-learning to students. Effective systems should be able to interpret learners' emotions and evaluate their attentiveness in order to deliver feedback that enhances their educational experiences.

However, developing these systems may be quite challenging. In emotion identification, it may be challenging to correlate emotional states to visual expressions. To overcome this, Paul Ekman attempted to map typical facial expressions

for emotions like contempt, fear, fury, sadness, and surprise [34]. In addition, late in the 1990s, when machines were able to identify emotion from both static photos and audio-visual input, 2 Wireless Communications and Mobile Computing drew further attention to this topic.

According to the literature, information about an individual's emotions may be gleaned by observing the face as a whole and paying close attention to the usage of the different facial muscles. [35] This is known as the sign-judgment strategy.

Using the Facial Action Coding System (FACS), facial expression action units (AUs) may be categorized and categorized according to emotion [36]. Automatic engagement recognition is another fascinating field. A real-time engagement recognition system might be used extensively in the following scenarios: (i) instructors working in distance education might get instant feedback based on their students' interest levels; (ii) participants' responses could be utilized to identify specific video segments. (iii) utilizes computer vision technologies that may evaluate student engagement in a distinct manner by examining body position, hand movements, and facial indications [37].

Learning, human intelligence, and emotion are all interconnected. Focus, learning motivation, and self-regulated learning are all impacted by emotions in learners. Through self-regulated learning and engagement, emotions, especially happy emotions, have a greater impact on academic performance. In e-learning, it is frequently seen that students become noisy during the same lectures or even courses as a result of unfavorable feelings. Additionally, associated learning material is activated in the long term memory by emotion. Positive feelings can thereby enhance students' ability to study more, perform well in assessments, and amass substantial knowledge. Numerous scientists have studied the identification of emotion in e-learning as a result of the connection between emotion and learning. Several scientists to investigate the e-ability learning's to recognize emotions.

A crucial aspect of every person is their emotional state, which affects their behavior, judgment, capacity for thought, adaptability, wellbeing, and interpersonal connections [38]. Emotions have a significant impact on human behavior, and human practices like e-learning must take this into account [39]. According to a study on the impact of experimentally induced positive and negative emotions on multimedia learning, students with the greatest previous knowledge or working skill could counterbalance the emotional influence on learning results.

According to [40, 41], e-learning promotes not just the learning process but also the connection between learning and emotion. As a consequence of the expansion of Learning Management Systems, traditional face-to-face learning is gradually being replaced by e-learning (LMS). Noteworthy is the significance of the data sources employed for emotion categorization. In typical classroom education, a teacher may alter his or her teaching style by analyzing students' facial expressions and body movements. However, this becomes difficult in e-learning situations.

According to study, a single data modality may not be able to capture the whole knowledge of the learning process. Therefore, several data sources are predicted [42, 43] to increase the accuracy of emotion classification [44]. EEG, EDA, eye tracking, audio, video, RB, and ECG are included in these data streams.

The authors of [45] also show the relevance of establishing robust user models and learning via the fusion of knowledge and technology. In reality, Learning Analytics and Knowledge (LAK) has recognized the significance of incorporating dynamic behavioral data in addition to traditional e-learning data (e.g., MOOCs, LMS data, etc.) [46]. Combining physiological data, such as electroencephalogram (EEG) or electrocardiogram (ECG), with external behaviors, such as eye movement or facial expressions, is a potential way for recording the sentiments and experiences of learners, according to [44]. According to the authors of [47], Multimodal Machine Learning (MML) is an approach for handling multimodal data sources as well as Data Harmonization of data [64].

Learning using diverse (multimodal) sources enables you to see how multiple modalities interact and provides a comprehensive understanding of how natural events operate. Recent research [48] indicates that incorporating multimodal data boosts accuracy and provides a better knowledge of the learner's emotions and experiences. Appendix A discusses the RQ1 in further detail.

RQ #2: How emotional and psychological states observed/examined in e-learning platforms?

This research question aims to identify how emotional and psychological states are measured by the techniques or algorithms used to measure them. Selected studies in this Research question show that the e-learning domain is used as it shows a high impact on higher education institutes. State types shown in Appendix B are both psychological and emotional as education institutes implemented it to understand e-learning's impact on students ethically and morally. The learning type mentioned in these studies is e-learning.

Whereas the techniques used to identify the states are diverse, such as AI, ML, DL, Signal system, mathematical representation. The techniques under AI, ML and DL are BiLSTM, Feed-Forward Neural Network, Hierarchical attention network, CNN, DESNet, SVM, Naïve Bayes, Logistic regression, Linear SVC and multinomial NB, Random Forests. Signal systems uses the concept of mathematical representations to measure the states. The techniques or methods are Gaussian mixing model (GMM), Partial least square structural equation modeling method, Gabor filter bank, MFCC features.

RQ #3: Which type of algorithms and techniques admitted for the emotional and psychological states in e-learning platforms?

Algorithm and techniques adopted by e-learning platforms for emotional and psychological states shown in Appendix C.

#### A. Contribution

To the best of my knowledge, this is the first SLR to discuss the emotional and psychological states as one unit, the

mathematical methods, and signal and AI-based techniques applied in e-learning platforms. The main scientific contribution of this SLR is that it will be helpful for the government to adopt measures for mental health by putting criteria for psychological and emotional states. As well as practically implemented in Higher Education Institutes to check teaching and learning performance. Also, this SLR focuses only on the AI and related tools which are widely used nowadays and help organizations to implement emotional and psychological state measures while teaching and conducting training through e-learning platforms.

Based on the findings and discussion, the information provided by this SLR will be helpful for researchers and stakeholders in applying these approaches and techniques, which deal with the wide variety of e-learning training and webinars. As previously said, the most current approaches for machine learning, deep learning, and neural networks would aid in retrieving, representing, and displaying recently used data.

#### B. Implication for Practice

This study has several practical implications on the provision of e-learning platform technology in higher education institutions and training institutes around the globe. It will help the government reduce the percentage of psychological and emotional pressure in youth and young children and helps in making work balance environment in all organizations in the country. Also, it will enhance the learning capabilities in students and teachers by adopting the latest tools and techniques that are mentioned in Appendix B and C.

#### C. Limitations

Only e-learning department was targeted, and the major focus was health and education. The studies are excluded related to learning dance, music, art, craft, sports and so on. The included studies are only that are written in English. The studies that only focuses on implementation of AI algorithms are included and all theoretical and conceptual models are excluded.

#### D. Future Suggestions

Emotional and psychological factors affect the public due to sudden changes in state and federal governing bodies. Emotional and psychological factors effects on public due to the retirement of favorite players from sports, politics, school, college or university. Emotional and psychological factors affect public due to inflation rate and change in prices in daily use items. Emotional and psychological factors effects on public due to political and shocking news related to interfaith harmony.

In addition, comparative research based on online, hybrid/blended formats are required to understand how the outcomes vary and how these changes impact the e-learning design framework. Comparative studies of the effectiveness of e-learning systems at various levels, such as the impact felt by learners vs. the effects experienced by instructors, are necessary.

## VI. CONCLUSION

Education may be improved, streamlined, and made more effective through an online learning platform. The development of e-learning has made it possible for students to get an education whenever and wherever they like. They make it possible to receive customized instruction and information at any time. These online instructional methods profoundly influence learners' mental and emotional states, affecting their skills and motivation. Distance between tutor and student, both in terms of space and time, makes it difficult to inspire students, who may experience a range of emotions from annoyance to uncertainty to ambivalence.

The student's motivation to learn, as well as his or her emotional and psychological well-being, will take a serious hit as a result, both during and after the class. This study aimed to investigate the techniques now employed by academic institutions for analyzing human sentiments and mental states. Only e-learning, alongside other fundamental technologies like machine learning, deep learning, signal processing, and mathematical techniques, has been employed in the field of education, as shown by the results.

These strategies are employed in a wide range of practical applications, emphasizing online education. The most common research techniques are analyzed, and each topic is presented in detail. In conclusion, we offer a detailed assessment of the state of the art, our contributions, and their implications, as well as our limitations and recommendations for future study.

## REFERENCES

- [1] Kumar Basak, Sujit, Marguerite Wotto, and Paul Belanger. "E-learning, M-learning and D-learning: Conceptual definition and comparative analysis." *E-learning and Digital Media* 15.4 (2018): 191-216.
- [2] Choudhury, Snigdha, and Snigdha Pattnaik. "Emerging themes in e-learning: A review from the stakeholders' perspective." *Computers & Education* 144 (2020): 103657.
- [3] Alqahtani, Ammar Y., and Albraa A. Rajkhan. "E-learning critical success factors during the covid-19 pandemic: A comprehensive analysis of e-learning managerial perspectives." *Education sciences* 10.9 (2020): 216.
- [4] Toth-Stub, S. "Countries Face an Online Education Learning Curve: The Coronavirus Pandemic has Pushed Education Systems: Online, Testing Countries' Abilities to Provide Quality Learning for All". 2020. Available online: <https://www.usnews.com/news/best-countries/articles/2020-04-02/coronaviruspandemic-tests-countries-abilities-to-create-effective-online-education> (accessed on 27 April 2020).
- [5] COVID-19 Educational Disruption and Response. 2020. Available online: <https://en.unesco.org/covid19/educationresponse> (accessed on 19 May 2020).
- [6] Settineri, Salvatore, et al. "Psychological types and learning styles." *Mediterranean Journal of Clinical Psychology* 6.3 (2018).
- [7] Landmann, Helen. "Emotions in the context of environmental protection: Theoretical considerations concerning emotion types, eliciting processes, and affect generalization." (2020).
- [8] Abbas, Z.; Umer, M.; Odeh, M.; McClatchey, R.; Ali, A.; Farooq, A. "A semantic grid-based e-learning framework (SELF)". In *Proceedings of the CCGrid 2005. IEEE International Symposium on Cluster Computing and the Grid 2005*, CWL, UK, 9–12 May 2005; Volume 1, pp. 11–18.
- [9] Muhammad, A.; Ghalib, M.F.M.D.; Ahmad, F.; Naveed, Q.N.; Shah, A. "A study to investigate state of ethical development in e-learning". *J. Adv. Comput. Sci. Appl.* 2016, 7, 284–290. [CrossRef].
- [10] Naveed, Q.N.; Muhammad, A.; Sanobar, S.; Qureshi, M.R.N.; Shah, "A. A mixed method study for investigating critical success factors (CSFs) of e-learning in Saudi Arabian universities". *Methods* 2017, 10. [CrossRef].
- [11] Hameed, S.; Badii, A.; Cullen, A.J. "Effective e-learning integration with traditional learning in a blended learning environment". In *Proceedings of the European and Mediterranean Conference on Information Systems*, Al Bustan Rotana, Dubai, UAE, 25–26 May 2008; pp. 25–26.
- [12] Basak, S.K.; Wotto, M.; Bélanger, P. "A framework on the critical success factors of e-learning implementation in higher education: A review of the literature". *Int. J. Educ. Pedagog. Sci.* 2016, 10, 2409–2414.
- [13] Al-Dosari, H. "Faculty members and students perceptions of e-learning in the English department: A project evaluation". *J. Soc. Sci.* 2011, 7, 291. [CrossRef].
- [14] Al-Asmari, A.M.; Khan, M.S.R. "E-learning in Saudi Arabia: Past, present and future". *Near Middle East. J. Res. Educ.* 2014, 2014, 2. [CrossRef].
- [15] <https://news.berkeley.edu/2017/09/06/27-emotions/#:~:text=The%2027%20emotions%3A%20admiration%2C%20adoration,%2C%20satisfacti on%2C%20sexual%20desire%2C%20surprise>.
- [16] Graham, C.R.; Woodfield, W.; Harrison, J.B. "A framework for institutional adoption and implementation of blended learning in higher education". *Internet High. Educ.* 2013, 18, 4–14. [CrossRef].
- [17] Mohammed, H.J.; Kasim, M.M.; Shaharane, I.N. "Evaluation of E-learning approaches using AHP-TOPSIS technique". *J. Telecommun. Electron. Comput. Eng. (JTEC)* 2018, 10, 7–10.
- [18] Dweiri, F.; Kumar, S.; Khan, S.A.; Jain, V. "Designing an integrated AHP based decision support system for supplier selection in automotive industry". *Expert Syst. Appl.* 2016, 62, 273–283. [CrossRef].
- [19] Anggrainingsih, R.; Umam, M.Z.; Setiadi, H. "Determining e-learning success factor in higher education based on user perspective using Fuzzy AHP". *MATEC Web Conf.* 2018, 154, 03011. [CrossRef].
- [20] Thai, N.T.T.; De Wever, B.; Valcke, M. "The impact of a flipped classroom design on learning performance in higher education: Looking for the best "blend" of lectures and guiding questions with feedback". *Comput. Educ.* 2017, 107, 113–126. [CrossRef].
- [21] Young, T.P.; Bailey, C.J.; Guptill, M.; Thorp, A.W.; Thomas, T.L. "The flipped classroom: A modality for mixed asynchronous and synchronous learning in a residency program". *West. J. Emerg. Med.* 2014, 15, 938. [CrossRef].
- [22] Alhabeeb, A.; Rowley, J. "E-learning critical success factors: Comparing perspectives from academic staff and students". *Comput. Educ.* 2018, 127, 1–12. [CrossRef].
- [23] Scholkmann, A. "What I learn is what I like. How do students in ICT-supported problem-based learning rate the quality of the learning experience, and how does it relate to the acquisition of competences?" *Educ. Inf. Technol.* 2017, 22, 2857–2870. [CrossRef].
- [24] Rowe, J.A. "Synchronous and Asynchronous Learning: How Online Supplemental Instruction Influences Academic Performance and Predicts Persistence". Ph.D. Thesis, Capella University, Minneapolis, MN, USA, 2019.
- [25] Abdel-Gawad, T.; Woollard, J. "Critical success factors for implementing classless e-learning systems in the Egyptian higher education". *Int. J. Instr. Technol. Distance Learn.* 2015, 12, 29–36.
- [26] Alhabeeb, A.; Rowley, J. "Critical success factors for eLearning in Saudi Arabian universities". *Int. J. Educ. Manag.* 2017, 31, 131–147. [CrossRef].
- [27] Bhuasiri, W.; Xaymoungkhoun, O.; Zo, H.; Rho, J.J.; Ciganek, A.P. "Critical success factors for e-learning in developing countries: A comparative analysis between ICT experts and faculty". *Comput. Educ.* 2012, 58, 843–855. [CrossRef].
- [28] Behzadian, M.; Otahsara, S.K.; Yazdani, M.; Ignatius, J. "A state-of-the-art survey of TOPSIS applications". *Expert Syst. Appl.* 2012, 39, 13051–13069. [CrossRef].
- [29] Muianga, X.; Klomsri, T.; Tedre, M.; Mutimucio, I. "From teacher-oriented to student-centred learning: Developing an ict-supported learning approach at the eduardo mondlane university, mozambique". *Turk. Online J. Educ. Technol.* 2018, 17, 46–54.

[30] Keele, S. "Guidelines for Performing Systematic Literature Reviews in Software Engineering"; Technical Report, Ver. 2.3 EBSE Technical Report; EBSE: Goyang-si, Korea, 2007.

[31] Kumar, Ganesh, et al. "Data harmonization for heterogeneous datasets: a systematic literature review." *Applied Sciences* 11.17 (2021): 8275.

[32] Pise, Anil Audumbar, Hima Vadapalli, and Ian Sanders. "Estimation of learning affects experienced by learners: an approach using relational reasoning and adaptive mapping." *Wireless Communications and Mobile Computing* 2022 (2022).

[33] S. ter Stal, L. L. Kramer, M. Tabak, H. op den Akker, and H. Hermens, "Design features of embodied conversational agents in health: a literature review," *International Journal of Human-Computer Studies*, vol. 138, Article ID 102409, 2020.

[34] J. B. Engelmann and M. Pogosyan, "Emotion perception across cultures: the role of cognitive mechanisms," *Frontiers in Psychology*, vol. 118, no. 4, 2013.

[35] S. L. Happy, A. Dasgupta, P. Patnaik, and A. Routray, "Automated alertness and emotion detection for empathic feedback during e-learning," in *Proceedings of the 2013 IEEE Fifth International Conference on Technology for Education (t4e 2013)*, pp. 47–50, IEEE, Kharapur, India, December 2013.

[36] T. Skiendziel, A. G. R'osch, and O. C. Schultheiss, "Assessing the convergent validity between the automated emotion recognition software noldus facereader 7 and facial action coding system scoring," *PLoS One*, vol. 14, no. 10, Article ID e0223905, 2019.

[37] D. Prakash, J. Van Haneghan, W. Blackwell, S. Jackson, G. Murugesan, and K. S. Tamilselvan, "Classroom engagement evaluation using computer vision techniques," in *Pattern Recognition and Tracking XXX*, M. S. Alam, Ed., vol. 10995, pp. 192–199, International Society for Optics and Photonics, Bellingham, DC, USA, 2019.

[38] Ekman, P.: "An argument for basic emotions". *Cognition and Emotion*6(3-4), 169–200 (1992).

[39] Faria, A.R., Almeida, A., Martins, C., Gonçaves, R., Martins, J., Branco, F.: "A global perspective on an emotional learning model proposal". *Telematics and Informatics*34(6), 824 – 837 (2017).

[40] Bahreini, K., Nadolski, R., Westera, W.: "Towards multimodal emotion recognition in e-learning environments". *Interactive Learning Env.*24(3), 590–605 (2016).

[41] Finch, D., Peacock, M., Lazdowski, D., Hwang, M.: "Managing emotions: A case study exploring the relationship between experiential learning, emotions, and student performance". *Int'l Journal of Management Education*13(1), 23–36 (2015).

[42] Blikstein, P., Worsley, M.: "Multimodal learning analytics and education data min-ing: Using computational technologies to measure complex learning tasks". *J. Learn. Anal.*3, 220–238 (09 2016).

[43] Prieto, L., Sharma, K., Kidzinski, L., Rodríguez-Triana, M., Dillenbourg, P.: "Multimodal teaching analytics: Automated extraction of orchestration graphs from wearable sensor data". *J. Comput. Assist. Learn.*34(2), 193–203 (2018).

[44] Zheng, W., Liu, W., Lu, Y., Lu, B., Cichocki, A.: "Emotionmeter: A multimodal framework for recognizing human emotions". *IEEE Transactions on Cybernetics*49(3), 1110–1122 (March 2019).

[45] Di Mitri, D., Scheffel, M., Drachler, H., Börner, D., Ternier, S., Specht, M.: "Learn-ing pulse: A machine learning approach for predicting performance in self-regulated learning using multimodal data". p. 188–197. *ACM* (2017).

[46] Mitri, D.D., Schneider, J., Specht, M., Drachler, H.: "The big five: Addressing recurrent multimodal learning data challenges". vol. 2163. *CrossMML* (2018).

[47] Baltrušaitis, T., Ahuja, C., Morency, L.: "Multimodal machine learning: A survey and taxonomy". *IEEE TPAMI*41(2), 423–443.

[48] Lee, D.H., Anderson, A.K.: "Reading what the mind thinks from how the eye sees". *Psychological Science*28(4), 494–503 (2017).

[49] Li, Jiachen, Songhua Xu, and Xueying Qin. "A hierarchical model for learning to understand head gesture videos." *Pattern Recognition* 121 (2022): 108256.

[50] BENNANI, Samir, and Mustapha BASSIRI. "Automatic Identification of E-Learner Emotional States to Ameliorate Her/His Motivation.".

[51] El-Ashry, A., El-Din, A. N., Khairy, K., Soliman, P., Salah, R., & Nosier, S. "Determining the Critical Success Factors (CSFs) Influencing E-learning in High Education, using the Partial Least Squares Structural Equation Modelling.

[52] Kouahla, Med Nadjib, et al. "Emorec: a new approach for detecting and improving the emotional state of learners in an e-learning environment." *Interactive Learning Environments* (2022): 1-19.

[53] Nandi, Arijit, et al. "Real-time multimodal emotion classification system in e-learning context." *Proceedings of the 22nd Engineering Applications of Neural Networks Conference: EANN 2021*. Cham: Springer International Publishing, 2021.

[54] Chanaa, Abdessamad. "E-learning Text Sentiment Classification Using Hierarchical Attention Network (HAN)." *International Journal of Emerging Technologies in Learning (IJET)* 16.13 (2021): 157-167.

[55] Hamal, Oussama, et al. "Boosting E-learner's Motivation through Identifying his/her Emotional States." *Iraqi Journal of Science* (2021): 127-132.

[56] Semerci, Yusuf Can, and Dionysis Goularas. "Evaluation of students' flow state in an e-learning environment through activity and performance using deep learning techniques." *Journal of Educational Computing Research* 59.5 (2021): 960-987.

[57] Karumuri, Sri Rama Murthy, V. Kamakshi Prasad, and Pavan Srinivas Narayana. "Emotion Detection of Students While Adopting E-Learning." Pavan Srinivas, *Emotion Detection of Students While Adopting E-Learning*.

[58] Alkhalaf, Salem, et al. "Emotional Intelligence Robotics to Motivate Interaction in E-Learning: An Algorithm." *International Journal of Advanced Computer Science and Applications* 12.6 (2021).

[59] Wang, Ruijie, Liming Chen, and Aladdin Ayesh. "Multimodal motivation modelling and computing towards motivationally intelligent E-learning systems." *CCF Transactions on Pervasive Computing and Interaction* (2022): 1-18.

[60] Du, Yu, Rubén González Crespo, and Oscar Sanjuán Martínez. "Human emotion recognition for enhanced performance evaluation in e-learning." *Progress in Artificial Intelligence* (2022): 1-13.

[61] Tseng, Fan Hsun, et al. "Real-time Facial Expression Recognition via Dense & Squeeze-and-Excitation Blocks." *Human-centric Computing and Information Sciences* 12 (2022): 39.

[62] Rajesh, P., and D. Akila. "Sentimental analysis on E-Learning videos using Hybrid Algorithm based on Naïve Bayes and SVM." *2022 International Conference on Emerging Smart Computing and Informatics (ESCI)*. IEEE, 2022.

[63] Ismail, Heba, et al. "Triggers and Tweets: Implicit Aspect-Based Sentiment and Emotion Analysis of Community Chatter Relevant to Education Post-COVID-19." *Big Data and Cognitive Computing* 6.3 (2022): 99.

[64] Kumar, Ganesh, et al. "Data Harmonization for Heterogeneous Datasets in Big Data-A Conceptual Model." *Software Engineering Perspectives in Intelligent Systems: Proceedings of 4th Computational Methods in Systems and Software 2020*, Vol. 1 4. Springer International Publishing, 2020.

APPENDIX A

Study	Domain	State Type	Learning type	Format
[49]	Health and Education	Psychological	E-learning	Video
[50]	Education	Emotional	E-Learning	Speech
[51]	Education	Psychological	E-Learning	Text

[52]	Education	Emotion and psychological	E-Learning	Speech
[53]	Education	Emotional	E-Learning	Multimodal
[54]	Education	Emotional	E-Learning	Image and Text
[55]	Education	Emotional	E-Learning	Image and Text
[56]	Education	Psychological	E-Learning	Image
[57]	Education	Emotion	E-Learning	Image
[58]	Education	Emotion	E-Learning	Image
[59]	Education	Psychological	E-Learning	Multimodality
[60]	Education	Emotional	E-Learning	Image
[61]	Education	Emotional	E-Learning	Image
[62]	Education	Emotional	E-Learning	Video
[63]	Education	Emotion and psychological	E-Learning	Tweets

APPENDIX B

Study	Domain	State Type	Learning type	Technique/Method
[49]	Health and Education	Psychological	E-Learning	BiLSTM
[50]	Education	Emotional	E-Learning	Gaussian mixing model
[51]	Education	Psychological	E-Learning	Partial least square structural equation modeling method
[52]	Education	Emotion and psychological	E-Learning	Gabor filter bank, MFCC features
[55]	Education	Emotional	E-Learning	Gaussian mixing model (GMM)
[53]	Education	Emotional	E-Learning	Feed-Forward Neural Network
[54]	Education	Emotional	E-Learning	Hierarchical attention network,
[56]	Education	Emotion	E-Learning	CNN
[58]	Education	Emotion	E-Learning	Proposed Algorithm
[60]	Education	Emotional	E-Learning	E-Learning Heuristic Multimodal
[61]	Education	Emotional	E-Learning	Dense Squeeze-and-Excitation Networks (DSENet)
[62]	Education	Emotional	E-Learning	SVM and Naïve Bayes algorithms are fused to be used as a Hybrid algorithm
[63]	Education	Emotion and psychological	E-Learning	Logistic Regression, Linear SVC Multinomial NB, Random Forests

APPENDIX C

Study	Domain	State Type	Learning type	Algorithm /Technique	Evaluation	Performance
[49]	Health and Education	Psychological	E-Learning	BiLSTM	They provide a new error threshold for this evaluation task. If the difference between the algorithm's evaluation and the gold standard label is less than a certain threshold, we accept the algorithm's assessment as accurate; otherwise, we reject the algorithm's assessment as wrong.	We also compute an error score for each assessment task, where an error is the absolute difference between an algorithmically assessed score and the corresponding score in the gold standard, to better examine the relative performance difference across these methods.
[50]	Education	Emotional	E-Learning	Supervised	Not mentioned	Therefore, we were able to capitalise on and collect data that could be used to gauge a student's cognitive health and inspire more engagement in the classroom.
[51]	Education	Psychological	E-Learning	Partial Least Square Model	The internal and external dependability of the model are tested using Cronbach's alpha, the Composite Reliability (CR) test, and the AVE method.	The technology factor, E-Learning factor, intention factor, and user satisfaction factor all have CR scores of 0.81, 0.88, and 0.64, respectively.
[52]	Education	Emotion and psychological	E-Learning	Proposed algorithm	Wilcoxon's one-sided matched test was utilised in the first experiment. For this second set of data, the Mann-Whitney U test was utilised.	When compared to the significance level, the estimated Pvalue is less than the significance threshold was set at the.05 alpha level. P-values below the significance level are not calculated. significance level = 0.05 (Pvalue = 0.00018 0.05).
[53]	Education	Emotional	E-Learning	Feed-Forward Neural Network,	Weighted Majority Voting	Valence: Accuracy 0.8477 F1-score 0.8649 Arousal: Accuracy 0.9551

				Incremental Stochastic Gradient Descent		F1-score 0.9589
[54]	Education	Emotional	E-Learning	Naive Bayes (NB), Logistic regression (LR), Support Vector Machine (SVM) and Multi-layer perception (MLP)	The threshold for training batches is 50 words or sentences, and each batch consists of 16 individuals. The stochastic gradient descent at a speed of 0.9 A 0.1 rate of learning was used for a total of 10 iterations.	HAN outperformed the other prediction results obtained from standard text classification algorithms with its 70.3% classification accuracy.
[55]	Education	Psychological	E-Learning	Convolutional autoencoder model architecture	Principal Component Analysis	The majority of students in both classes are engaged and participating in their learning, as depicted by the flow charts. For this finding to hold, it is necessary to assume that the course material was satisfactory to the vast majority of participants.
[56]	Education	Emotion	E-Learning	CNN	In order to move a point from the Image domain (top panel) to the Hough transform domain, the Hough transform is applied (bottom panel).	The classifier is developed using 70% of the data, and its predictive accuracy is evaluated using 30% of the original data.
[57]	Education	Emotion	E-Learning	<i>Emotional web assistance for EREIL</i>	When a student is communicating in the classroom, EREIL can read cues from the student's body (such as emotions, volume of voice, gestures, etc.)	EREIL is able to learn about a student's nonverbal cues through interacting with their eyes, gestures, facial analysis, and voice recognition. This newfound knowledge allows EREIL to positively identify students.
[58]	Education	Psychological	E-Learning	Logistic regression, Open Gaze And Mouse Analyzer (OGAMA) 5.0	Executed on the basis of the motivation evaluation; deducing the Each of the inspiring factors can be ranked on a high or low scale, machine learning factors using logistic regression classifier.	The precision of The best threshold for EEG and eye tracking-based motivating factor prediction ranges from 68.1% to 92.8% values.
[59]	Education	Emotion and psychological	E-Learning	Empirical research method i.e V.E. Milman	seven The several varieties of motivational outlooks are discussed. of mental effort devoted to informational actions. Using this method, consisting of seven distinct measures of intrinsic motivation, and being of a standard profile	Not mentioned
[60]	Education	Emotional	E-Learning	navy Bayesian classifier networks	The superiority of a dynamic system for the rational mind, implying that emotional data could considerably improve the effectiveness of the e-learning platform.	The end result is an e-learning success rate of 93.85%, a hand gesture success rate of 92.70%, a speech recognition success rate of 82.26%, a decrease in emotional problem success rates of 84.50%, and so on.
[61]	Education	Emotional	E-Learning	Dense Squeeze-and-Excitation Networks (DSENet)	The Facial Expression Recognition 2013 open dataset is used for DSENet's training and validation.	The model outperforms ResNet-34 by 6% when it comes to identifying emotional states.
[62]	Education	Emotional	E-Learning	SVM and Naïve Bayes algorithms are fused to be used as a Hybrid algorithm	Regression	The proposed hybrid approach achieves an accuracy of almost 97%.
[63]	Education	Emotion and psychological	E-Learning	Logistic Regression Linear SVC Multinomial NB Random Forests	To help decision-makers and staff in the educational sector improve and adjust the educational process during and after the pandemic, the created analytics are then factored by location and time to provide more thorough insights.	Linear Support Vector Classifier (SVC) performed best on all measures of accuracy, precision, recall, and F-measure (91%), according to a study of 11 different classifiers for emotions.

# Privacy and Integrity Verification Model with Decentralized ABE and Double Encryption Storage Scheme

Amrutha Muralidharan Nair<sup>1</sup>, Dr. R Santhosh<sup>2</sup>

Research Scholar, Department of Computer Science Engineering, Karpagam Academy of Higher Education, Coimbatore, India<sup>1</sup>  
Professor, Department of Computer Science Engineering, Karpagam Academy of Higher Education, Coimbatore, India<sup>2</sup>

**Abstract**—To support a wide range of applications, cloud computing has a variety of services. It has a number of positive acceptance tales as well as a couple of negative ones including security breaches. The versatile usage of cloud services to store sensitive and personal data in cloud become hesitated by many organizations because of security issues. A new model of relying on a third-party auditor (TPA) has been adopted to improve trust and entice adoption between cloud clients (CC). Hence, we require a dynamic approach to control the privacy and integrity problem that occur across the cloud computing. Decentralized Attribute based encryption techniques and FHE approach is used to overwhelmed the issues. In this proposed scheme, the integrity checking is verified and auditor by the TPA without have any knowledge of the data content and double encryption is performed on the data stored in cloud. the data owner encrypts the data using ABK-XE (Attribute Based Key generation with XOR encryption) technique and send it to tag server whose encrypt the data again using ECEA (Elliptical Curve Elgamal) algorithm and generate the signature and unique ID using SHA-1 algorithm then store the data in Cloud Environment. The proposed algorithm is an integration of auditing scheme with Symmetric key Encryption and Homomorphic Encryption.

**Keywords**—Cloud computing; data integrity; ABK-XE; ECEA Algorithm; SHAI

## I. INTRODUCTION

Cloud computing [1] is a ground-breaking computing model that has caught the interest of individuals from various fields, including academia and industry. Cloud computing is a large-scale distributed computing paradigm driven by economies of scale in which a pool of abstracted, virtualized, constantly increasing, managed computing power, storage, platforms, and services is supplied on demand to external clients through the Internet [2].

The fundamental component of the cloud is cloud storage, which directs the user's attention to data storage. However, the cloud server is not completely trustworthy, therefore the user confronts several security risks and problems while storing data in the cloud [3]. Data integrity is among the most serious privacy concerns since data hosted on cloud servers is not physically possessed by users and they have no control over it. While some people are prepared to give up their privacy in exchange for the benefits of software services, businesses and governments will never do so [4]. The integrity checking technique provides an efficient approach to examine the data

integrity on a cloud server. The client sends an integrity challenge to the cloud server, which creates a proof of the original data. If the evidence passes the verification procedure, the data is shown to be complete. Although numerous systems have been developed to assist data owners in ensuring the integrity of their data. Nonetheless, there are certain difficulties to be addressed. When the integrity verification is taking place, the public verifier is generally interested in the client data and attempts to get it. It is extremely risky for the owner to keep the data secret in light of the aforementioned issues. Encryption of the data is the simple way to make data secure but this method is not at all a good choice for the data user.

To solve the aforementioned issue, this work proposes an effective double encryption technique, and the verification process is done without knowledge of the original data, i.e., it is hidden from the public verifier. The contribution to the paper is listed:

- 1) A thorough understanding about the elliptical curve encryption.
- 2) A privacy preservation system with two tier encryption methodology.
- 3) Detail study about the existing system and usage of bilinear mapping, signing and verifying system.

The remainder of the paper is organized as follows. Section 2 gives a synopsis of the related research. The technique and mathematical formula are described in Section 3. Section 4: Explain the proposed system and its details. Section 5 examines and compares the performance of our system to that of other plans. Section 6 finishes with conclusions.

## II. RELATED WORK

Juels et al. [5] suggested the idea of “proof of retrievability (PoRs)”. This system not only validates the integrity of data saved in the cloud, but it also assures data retrievability through the use of error-correcting code. Nonetheless, it Private audit is performed.

Yu et al. [6] introduced the ID-based RDIC and its security architecture, which featured protection against a rogue cloud server and privacy from an external auditor with zero knowledge. The RDIC protocol does not leave information about the stored data auditor during the DRIC operation. In the generic group template, the new structure is proven to be

secure against a rogue server and achieves zero knowledge confidentiality against an auditor.

Ateniese et al. [7] developed a “Provable data Possession” (PDP) paradigm to achieve data integrity. The author [8] presented Another scheme scalable PDP scheme, which allows for block adding, updating, and deletion.

Sasikala et al. [9] presented a Remote Data Integrity Checking (RDIC) is critical for developing safe cloud storage. It allows users to validate the integrity of outsourced data without downloading the full file. The author also evaluated and assessed current RDIC techniques based on factors such as integrity testing method, cryptographic model, auditing mode, and data recovery. Finally, it sheds insight on unresolved topics such as research directions in the design of the RDIC procedure.

Yong et al. [10] presented an attribute-based cloud information integrity auditing protocol to reduce key management difficulties. The suggested technique requires significantly less calculation in validating the auditing reaction, resulting in reduced time consumption.

Yannan Li et al. [11] explored the difficult key management problem in cloud data integrity checking by presenting fuzzy identity-based auditing. Author introduced the concept of fuzzy identity-based data auditing, in which a user's identity may be regarded as a set of descriptive properties.

Feng et al. [13] provided a public remote integrity checking technique that protects user identification. This scheme, however, only supports file-level integrity verification.

Yu et al. [14] presented a novel identity-based public RDIC system that protects data privacy. However, the integrity testing of this technique necessitated a significant computational cost.

Tong et al. [15] devised a technique that delivers indistinguishable privacy (IND-privacy) as compared to TPA for both data content and timestamp. It build the authenticator with the randomizable structure-preserving signature to connect the content and timestamp in the authenticator and allow efficient timestamp updating (SPS). Furthermore, in the auditing phase, they use the Groth-Sahai proof and range proof to offer IND-privacy and ensure timestamp validity.

Zhang et al. [16] developed an RDIC technique that uses indistinguishability obfuscation to preserve data privacy while improving performance. However, this method is rigid and difficult to implement in practice.

Chen et al. [17] introduced a public verification technique based on algebraic signatures that employed a short bit string compressed by a data block to accomplish efficient integrity verification without comparing to original data. Unfortunately, it is vulnerable to a replay attack. It also does not enable dynamic data update.

### III. METHODOLOGY

It is necessary to develop a powerful audit model that solves the issue related to the existing system. The proposed

model is made in such a way that the cloud user (CU) can stored the data without heisting, since the data is stored in encrypted form (Fig. 2). The CU can send request to verify the data integrity through TPA. The TPA only send the request and accept the response from the proof-server (PS) without having a computational overhead. Also, it will not provide any information related to use file that is stored in cloud database (CS).

#### A. Preliminaries

Some of the basic Mathematical approach used in the auditing schemes and in the system model.

1) *Bilinear maps*: A bilinear mapping  $e: G_1 \times G_2 \rightarrow G_t$  where  $G_t$  is the multiplicative cyclic group of prime  $p$  with following properties.

- a)  $e$  is bilinear for all  $a, b \in Z_p$
- b)  $e$  is non-degenerate
- c)  $e$  is efficiently computable.

2) *Elliptic curve Elgamal Algorithm (ECEA)*: The Elliptical curve [12] with Elgamal algorithm converts the plaintext  $M$  to a point  $P_m$  on the elliptical curve  $E$ . The arithmetic operation on elliptical curve is as follows:

a) *Addition of two points*: Suppose  $A = \{X_a, Y_a\}$  and  $B = \{X_b, Y_b\}$ , where  $A \neq B$  that lie on elliptical curve  $e$ . the sum of  $A + B$  reults a third point  $C \{X_c, Y_c\}$  as shown in Fig. 1.

b) *Double pointing*: Let  $A = \{X_a, Y_a\}$  be a point lies on  $e$  adding the point  $A$  to itself is called Double pointing.

$$P + P = 2P$$

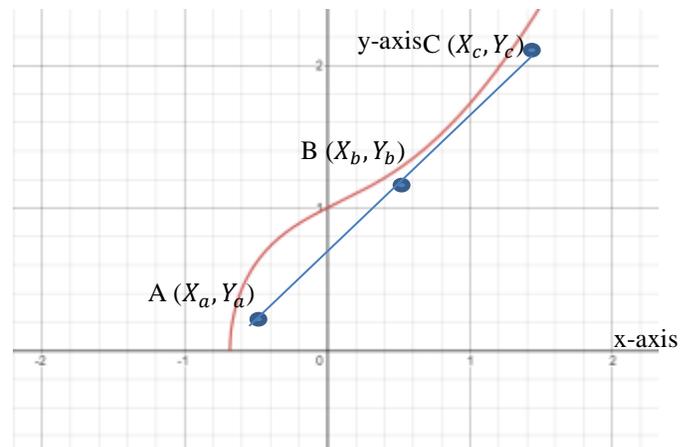


Fig. 1. Elliptical curve of  $A + B = C$ .

c) *Multiplication*: Suppose  $k$  is an integer  $A$  is a point  $(X_a, Y_a)$  then,

$$K_p = \frac{A + \dots + A}{K \text{ times}}$$

So, in this proposed scheme, the computation speed on the ECEA is reduced by using decimal representation of message. So that double pointing and addition operation can be avoided. This scheme involves the following procedures as; Suppose user  $A$  and  $B$  wishes to communicate with each other. The common parameter known by both users will be  $p$  and  $G$ , where  $p$  is the prime number and  $G$  is the base point in the

elliptical curve. Let user A and B choose their private key  $\alpha_a$  and  $\alpha_b$  over an interval  $[1, p-1]$  and generate the public key  $\beta_a = \alpha_a G$  and  $\beta_b = \alpha_b G$ .

The message or file M is divided into n block  $m_1, m_2, \dots, m_n$  which is then convert into decimal values  $d_1, d_2, \dots, d_n$  respectively such that the multiplication of basepoint and the decimal point transform to a point in E as:

$$\begin{aligned} p_{d1} &= d_1 G \\ p_{d2} &= d_2 G \\ &\vdots \\ p_{dn} &= d_n G \end{aligned}$$

Then user A computes the secret key K by multiplying with private key  $\alpha_a$  and B's Public key  $\beta_b$  as:

$$K = \alpha_a * \beta_b$$

This k value is added with the Decimal elliptical point to obtain cipher text as:

$$\begin{aligned} C_1 &= p_{d1} + K \\ C_2 &= p_{d2} + K \\ &\vdots \\ &\vdots \\ &\vdots \end{aligned}$$

$$C_n = p_{dn} + K$$

Therefore, the cipher text  $C = \{C_1, C_2, \dots, C_n\}$ . The plain text is obtaining by subtracting the secret key K as:

$K' = \alpha_b * \beta_a$  performing the following operation as:

$$\begin{aligned} p_{d1} &= C_1 - K' \\ p_{d2} &= C_2 - K' \\ &\vdots \\ &\vdots \\ p_{dn} &= C_n - K' \end{aligned}$$

Thus, obtain decimal values is converted to original values  $m_1, m_2, \dots, m_n$ .

**Proof,**

$$\begin{aligned} C_1 - K' &= C_1 - \alpha_b * \beta_a && (\because K' = \alpha_b * \beta_a) \\ &\Rightarrow p_{d1} + K - \alpha_b * \beta_a && (\because C_1 = p_{d1} + K) \\ &\Rightarrow p_{d1} + \alpha_a * \beta_b - \alpha_b * \beta_a && (\because K = \alpha_a * \beta_b) \\ &\Rightarrow p_{d1} + \alpha_a * \alpha_b G - \alpha_b * \alpha_b G \\ &&& (\because \beta_a = \alpha_a G \\ &&& \beta_b = \alpha_b G) \\ &\Rightarrow p_{d1} + \alpha_a * \alpha_b G - \alpha_b * \alpha_b G \\ &\Rightarrow p_{d1} \end{aligned}$$

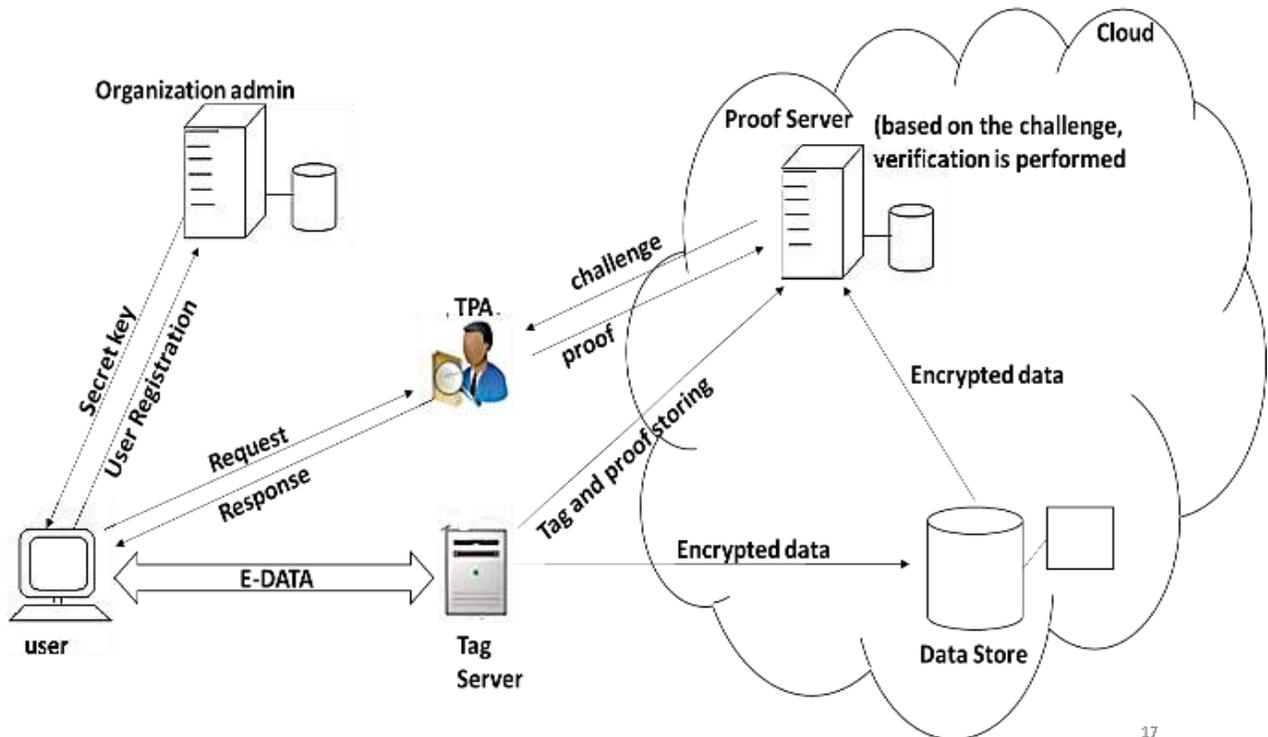


Fig. 2. Proposed system.

#### IV. PROPOSED SCHEME

The proposed scheme shows a threshold hybrid combination of encryption; this scheme is divided into CU encryption. TagServer processing, PS verification, TPA challenges.

1) *CU encryption*: In this cloud User (CU) encryption the file or data that he/she wishes to store in cloud by using its attribute UID and PW. This encryption/decryption is divided into two phases:

a) *Key generation process*: This generates a symmetric key based on user attributes:

- i. User will provide its password as the seed value, that will be convert into byte form as:

$$\text{byte } b[] = \text{byte}(\text{PW})$$

- ii. Calculate the sum as:

$$B = \sum_{i=0}^{b.length} b(i)$$

- iii. Applying log function and convert the value to a factor of 64

$$k_b = \log B$$

- iv. Using the bilinear paring the  $b(i)$  will be multiplied by  $k_b$  so that we obtain the symmetric key array  $A[i]$ ,

for each  $i = 0$  to  $b.length$

$$A[i] = k_b * b(i) \text{ mod } 64$$

end

b) *Encryption process*: The generated key  $A[i]$  will be XoR with the file  $F = \{f_1, f_2, \dots, \dots, f_n\}$  as follows.

- i. The file  $F$  is divided into  $n$  blocks of size 64 bits as  $= \{f_1, f_2, \dots, \dots, f_n\}$ , the last is padded with zeros if required.
- ii. Each block is XoR with the symmetric key as:

$$C[i] = \sum_{i=1}^n f[i] * A[i]$$

- iii. This obtain  $C(i)$  will be combined together to form the encrypted file  $E_A(F)$

The above algorithm is shown in Fig. 3.

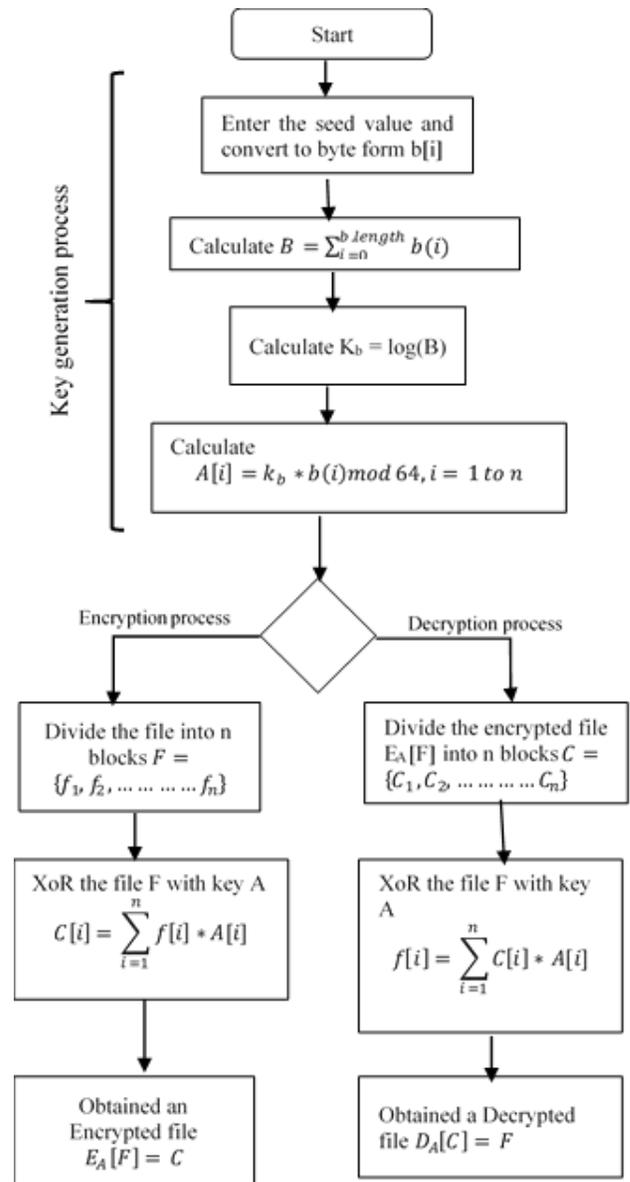


Fig. 3. User side encryption and decryption process.

2) *TagServer Processing (TSP)*: TSP plays a curial role in the scheme. Here the encrypted file  $E_A[F]$  received from the user undergoes ECEA encryption process and obtain the encrypted form as  $C^* = \text{Encrypt}_{ECEA}(E_A[F])$ . Then TSP calculates the Signature of the encrypted file  $C^*$  by using SHA-1 algorithm. Thus, signature digest is of the form 160-bit represented as  $Sig_{C^*} = MD(C^*)$ . TSP transfers the encrypted file  $C^*$  to cloud data store; also store the signature along with the unique file Id  $(Sig_{C^*} | F_{id})$ . TSP also send its public key and file Id to cloud user so that the CU can decrypt the file directly from the data store  $(F_{id} | \beta_b | T_1)$  as shown in Fig. 4.

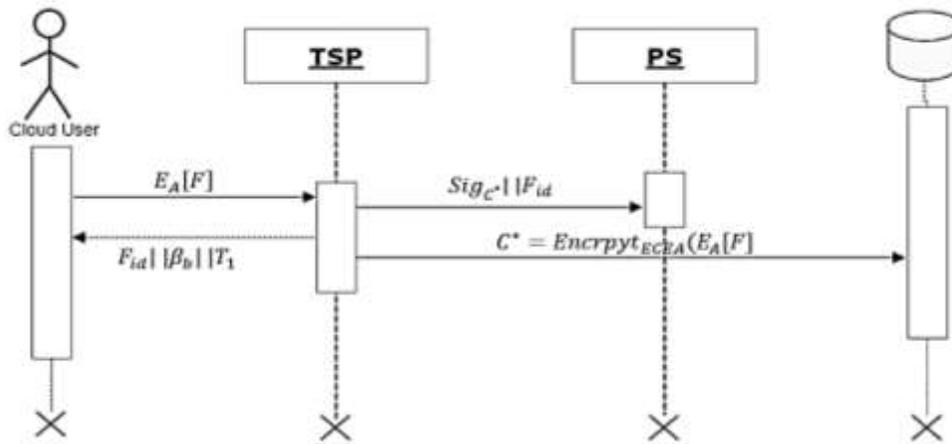


Fig. 4. Working of TSP.

3) *TPA challenges and Proof Server verification*: When the cloud user CU require to verify the file, it sends a challenge request to TPA as  $ChalR_Q = F_{id} || ID_{cu} || T_2$ . The auditing of the file is performed by sending a request to the TPA by the cloud user CU. The receive request  $ChalR_Q$  will be transmitted to proof Server PS.

The PS Retrieve the file from the data store using the  $F_{id}$ , compute the signature of the file  $Sig_{C^*}$  using SHA-1 Algorithm

and compare the signature store by TSP ( $Sig_{C^*}$ ) and new compute signature  $Sig_{C^*}^1$ . This comparison declared that the file stored in the data store is altered or not. This response  $Res_Q$  will be intimated to TPA which then forward to CU with the assurance of it file integrity is maintained.

Here the TPA and PS doesn't have direct access to the original file F so the semi trust issues is solved by using this scheme also it provides high assurance of data integrity in the cloud environment as shown in Fig. 5.

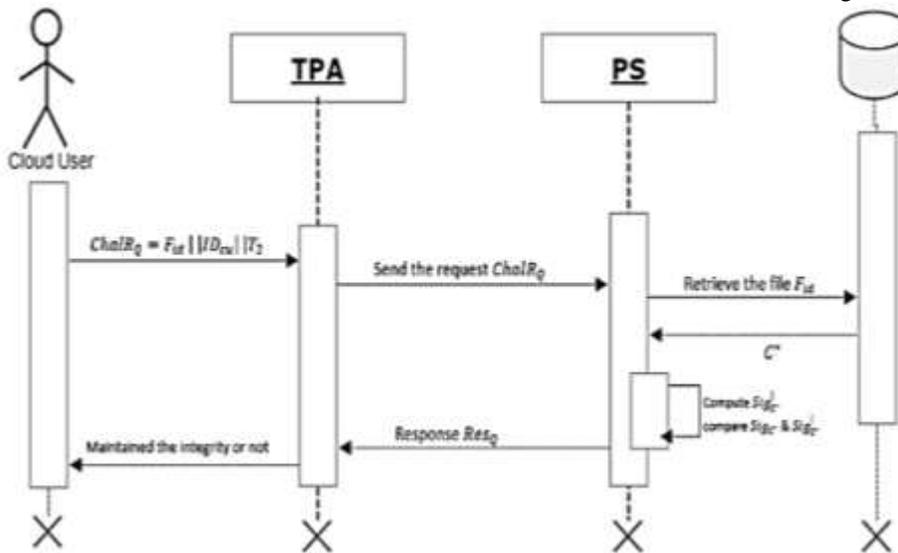


Fig. 5. Working process of TPA and PS.

## V. PERFORMANCE EVALUATION

The proposed scheme performance is summarized as follows.

### A. Computation Cost

The above specified algorithm contains operations such as multiplication, paring operation, addition and subtraction operation with the time  $T_m, T_p, T_a, T_s$  on the curve G respectively. The Time taken for other operation such as hash function and logarithm is were less so the computation cost for such operation can be negligible cost. Suppose the cloud user

contain  $m$  blocks in total,  $cl$  challenge is passed to TPA for checking the integrity of data block in cloud. The computation cost in TagServer algorithm for encrypting to run  $m$  times to encrypted the  $m$  blocks and generated the signature is  $(T_m(m + 1)) + mT_a$ . The computation cost at TPA to transfer the challenge is negligible cost. The computation cost at the proof Server is  $(cl + 1)T_m + clT_a$ . The verification cost of the proposed scheme is  $T_p + (cl + 1)T_m + clT_a$ . The decryption algorithm computation cost is  $(T_m(m + 1)) + mT_s$ . The below Table I shows the computation cost of our proposed model with Scheme [13] and [15].

TABLE I COMPUTATION COST COMPARISON

COST	SCHEMES		
	[13]	[15]	Proposed model
Encryption cost	$2T_{ex} + T_m$	$2T_{exp} + T_m$	$(T_m(m + 1)) + mT_a$
Proofgeneration	$2T_p + (cl + 1)T_{ex} + clT_m$	$(cl + 1)T_{ex} + (cll - 1)T_m$	$(cl + 1)T_m + clT_a$
Verification process	$clT_p + (cl + 1)T_{ex}$	$2T_p + (cl + 2)T_{ex} + (cl + 1)T_m$	$T_p + (cl + 1)T_m + clT_a$
Decryption process	$2T_{ex} - 1 + T_m$	$2T_{exp} - 1 + T_m$	$(T_m(m + 1)) + mT_s$

**B. Experimental Results**

To evaluate the efficiency of the proposed model, the model and compared model is implemented in Intel i5 core with windows 10 operating system, 8GB RAM and 1TBB hard disk. All the experiment was carried out by using different User of cloud environment for which the minimum configuration of 1 CPU is to make the research cost effective. It is simulated using CloudSim with NetBeans framework (Java language).

The main section of the proposed system is the encryption of encrypted file and generation of signature done by the tag server; verification of data integrity performs by the proof server. This scheme is evaluated by considering variable size block from 1KB to 100KB, as a result its encryption time, siggeneration time, sigverification time is obtained in millisecond(ms) as shown in the Table II. This result show that as the block size increases the encryption, signature generation or verifying time is less than the time taken for the 1 KB block. The result of the proposed scheme is shown in Fig. 6 and 7.

The efficiency of algorithm used for proof generation and verification is evaluated. Furthermore, we implemented the scheme [13] and [15] under the same experiment setting and make an efficiency comparison with each other. In the experiment the number of challenges in the system is increased. Thus, the result which is illustrated in Fig. 8 shows the efficiency of the algorithm when the challenges has increased from 100- 1000. Fig. 8 shows our scheme is little more efficient than the scheme [13] and [15] in proof generation. From Fig. 9, the verification purpose our scheme requires 5 second for 1000 challenge block and scheme [15] and [13] require almost 6.1 second for the same block.

TABLE II PROPOSED SYSTEM TIME COMPLEXITIES

Block size	TIME		
	Encrypting time	Proof generation	Proof verification
1 KB	185417	1138	1690
10 KB	183250	938	1234
20 KB	182189	930	1162
30 KB	180123	696	1088
40 KB	174960	464	1486
50 KB	152192	264	1424
60 KB	150215	440	1386
70 KB	147356	252	1028
80 KB	142523	232	1296
90 KB	135661	212	388
100 KB	128114	212	258

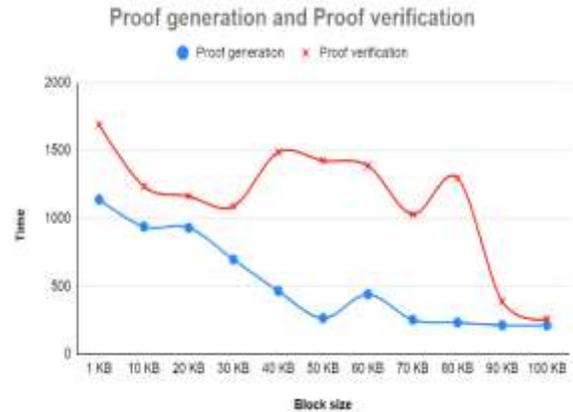


Fig. 6. Proof generation and verification time consumption.

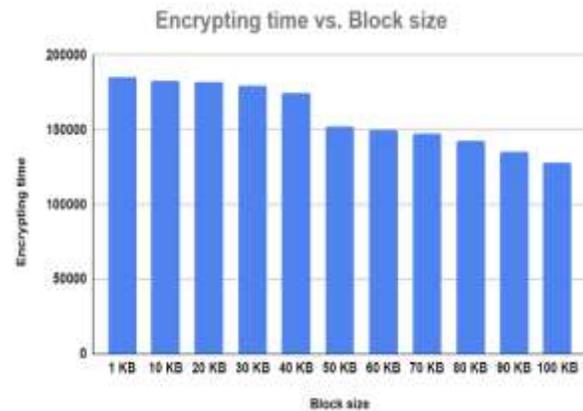


Fig. 7. Time consumption for encryption.

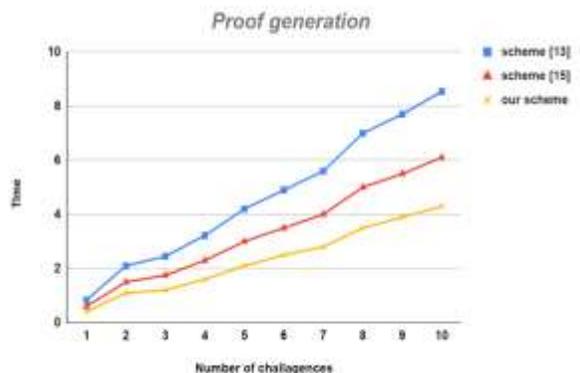


Fig. 8. Time consumption in proof generation process.

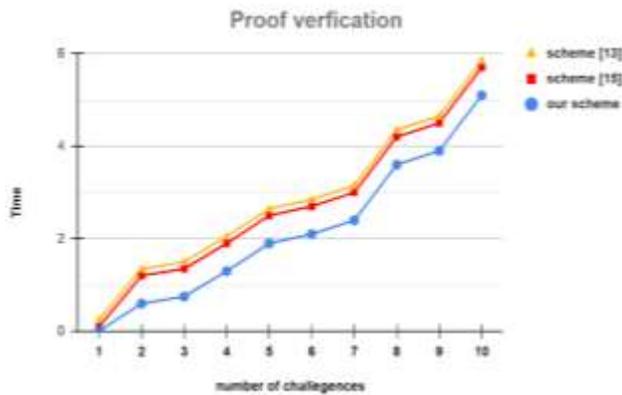


Fig. 9. Time consumption in verification process.

## VI. CONCLUSION

In cloud environment, data privacy is very much essential one. In this paper, we have proposed a technique which is a combination of symmetric and asymmetric encryption. The proposed scheme is a combination of XoR encryption and ECEA encryption with SHA-1 algorithm, which is suitable for integrity auditing of data stored in cloud computing. The existing schemes are meant to provide integrity maintenance for numerous data stored in cloud, but it doesn't provide dynamic data operation, data are visible to TPA totally.

The proposed scheme preserved the confidentiality of the file attributes and simplified the key management system in the traditional cloud facts. This scheme efficiently and effectively supports auditing tasks such as guaranteeing the TPA integrity, secure storage, overall control in the system. The result of the implementation which is shown above defines that the proposed scheme provides 0.1% more security than the existing protocols. This paper explains the complete details about the construction, implementation and experimental results of the proposed model.

## REFERENCES

[1] R. Buyya, C.S. Yeo, S. Venugopal, J. Broberg, and I. Brandic, "Cloud Computing and Emerging IT Platform: Vision, Hype, and Reality for Delivering Computing as the %th utility," *Future Gener. Comp. Syst.*, vol. 25, no. 6, pp. 599-616, 2009.

[2] Jiguo Li, Hao Yan, Yichen Zhang, "Certificate less public integrity checking of group shared data on cloud storage," *IEEE Transactions on Services Computing*, 2018.

[3] M. Ali, S.U. Khan, A.V. Vasilakos, "Security in Cloud Computing: Opportunities and Challenges," *Inf. Sci.*, vol. 305, no. 1, pp. 357-383, 2015.

[4] S. Suganya, "Improving Cloud Security by Enhancing Remote Data Integrity Checking Algorithm," *Innovations in Power and Advanced Computing Technologies (i-PACT) IEEE*, 2017.

[5] A. Juels and B. J. Kaliski, "PORS: Proofs of Retrieval for Large Files," in *In Proceedings of the 14th ACM Conference on Computer and Communications Security*, Alexandria, 2007.

[6] Yong Yu, Man Ho Au, Giuseppe Ateniese, Xinyi Huang, Willy Susilo, "Identity-Based Remote Data Integrity Checking With Perfect Data Privacy Preserving for Cloud Storage," *IEEE Transactions on Information Forensics and Security*, vol. 12, no. 4, pp. 767-778, 2017.

[7] Ateniese, Giuseppe & Burns, Randal & Curtmola, Reza & Herring, Joseph & Kissner, Lea & Peterson, Zachary & Song, Dawn, "Provable Data Possession at Untrusted Stores," in *Proceedings of the ACM Conference on Computer and Communications Security*, 2007.

[8] G. Ateniese, R.D. Pietro, L.V. Mancini, and G. Tsudik, "Scalable and Efficient Provable Data Possession," in *Proc. Fourth Int'l Conf. Security and Privacy in Commun. Netw. (SecureComm'08)*, 2008.

[9] Bindu, C. Sasikala and C. S., "A study on remote data integrity checking techniques in cloud," in *International Conference on Public Key Infrastructure and its Applications (PKIA)*, 2017.

[10] Yong Yu, Yannan Li, Bo Yang, Willy Susilo, Guoming Yang and Jian Bai, "Attribute-Based Cloud Data Integrity Auditing for Outsourced Storage." *IEEE Transaction on Emerging Topics in Computing* vol. 8, no. 2, pp. 377-390, 2017.

[11] Y. Li, Y. Yu, G. Min, W. Susilo, J. Ni and K. R. Choo, "Fuzzy Identity-Based Data Integrity Auditing for Reliable Cloud Storage Systems," *IEEE Transactions on Dependable and Secure Computing*, vol. 16, no. 1, pp. 72-83, 2019.

[12] A. Odlyzko, "Discrete Logarithms in Finite Fields and Their Cryptographic Significance," In *Workshop on the Theory and Application of Cryptographic Techniques*; Springer: Berlin/Heidelberg, Germany, 1987.

[13] Y. Feng, G. Yang and J. K. Liu, "A new public remote integrity checking scheme with user and data privacy," *International Journal of Applied Cryptography (IJACT)*, vol. 3, no. 3, pp. 196 - 209, 2017.

[14] Yong Yu, Man Ho Au, Giuseppe Ateniese, Xinyi Huang, Willy Susilo. "Identity-Based Remote Data Integrity Checking With Perfect Data Privacy Preserving for Cloud Storage." *IEEE Transactions on Information Forensics and Security*, vol.12, no. 4, pp: 767-778, 2017.

[15] Tong Wu, Guomin Yang, Yi Mu, Rongmao Chen, Shengmin Xu, "Privacy-enhanced remote data integrity checking with updatable timestamp", *Information Sciences*, vol. 527, pp:210-226, 2020.

[16] Y. Zhang, C. Xu, X. Liang, H. Li, Y. Mu and X. Zhang, "Efficient Public Verification of Data Integrity for Cloud Storage Systems from Indistinguishability Obfuscation," in *IEEE Transactions on Information Forensics and Security*, vol. 12, no. 3, pp. 676-688, March 2017.

[17] Chen, L. Using Algebraic Signatures to Check Data Possession in Cloud Storage. *Future Gener. Comput. Syst.*, vol. 29, pp:1709-1715, 2013.

# Design of a Hybrid Recommendation Algorithm based on Multi-objective Collaborative Filtering for Massive Cloud Data

Xiaoli Zhou

Sichuan Top IT Vocational Institute, School of Information Engineering, Chengdu, 610000, China

**Abstract**—The current recommendation technology has some problems, such as lack of timeliness, the contradiction between recommendation diversity and accuracy. In order to solve the problem of lack of timeliness, the time factor is introduced when constructing the self-preference model. The cold start problem in the collaborative filtering algorithm is solved by the hybrid similarity calculation method, and the potential preference model is constructed. The two are fused to obtain a hybrid recommendation algorithm to improve the recommendation performance of the algorithm. For the problem of multi-objective contradiction, the NNIA algorithm is used to further optimize the candidate results of mixed recommendation, and the final recommendation list is obtained. Through verification experiments, the results show that the recall rate and accuracy of the fused preference model are better than those of the non-fused model, and the accuracy is 9.57% and 8.23% higher than that of SPM and PPM, and the recall rate is 9.97% and 7.65% higher, respectively. CBCF-NNIA algorithm has high accuracy and diversity of recommendation, and can provide users with rich and diverse text content to meet their own needs.

**Keywords**—Self preference; collaborative filtering; potential preferences; mixed recommendation; multi-objective optimization

## I. INTRODUCTION

With the advent of the information age, data has become a decisive factor in the development of the industry, and any decision needs to rely on data to speak. In the face of massive data resources, storing them in the cloud to form cloud data that is easy to manage is beneficial for users to access relevant information [1]. The constant development of current Internet technology and communication technology has led to a rapid increase in the amount and speed of information and data dissemination, making it impossible for users to precisely find useful information and making a large amount of information unavailable, which leads to information overload [2]. The personalized recommendation system can analyze the user's interest preferences according to the user's behavior habits, and recommend the content related to the user's interest to the user without the user's initiative to provide information. The excellent performance makes the recommendation system widely used in many fields [3-4]. At present, there are many commonly used recommendation methods. Collaborative Filtering (CF) recommendation is the earliest recommendation method. However, with the growth of data size, data sparsity, system could start and other problems are unavoidable. The hybrid recommendation algorithm can integrate multiple recommendation algorithms, and research the combination of

collaborative filtering and content recommendation method to form a cascade hybrid recommendation algorithm. It is expected that the hybrid recommendation algorithm can effectively alleviate the cold start problem of the system and improve the accuracy of the recommendation results. In view of the contradiction between the diversity and accuracy of the recommendation results, the multi-objective immune optimization algorithm is studied and introduced to further optimize the recommendation list. It is expected that this method can find the best recommendation results, which can achieve the diversity of the results while ensuring the accuracy of the recommendation, and meet the multiple needs of users.

## II. RELATED WORK

The research team of Tian proposes a book recommendation system based on a hybrid recommendation algorithm to address the problem of users finding appropriate books quickly. The system combines collaborative filtering with content recommendation algorithms, improves the user item rating matrix, and uses clustering algorithms to solve the data sparsity problem. Through practical application, the results show that the hybrid recommendation method studied can provide users with more accurate book recommendations [5]. Wang scholars have studied a hybrid recommendation algorithm based on interest models in order to improve user satisfaction on e-commerce websites. The algorithm uses collaborative filtering to mine users' potential interests and a content recommendation algorithm to construct a model of users' existing interests, and the two are combined to recommend highly accurate and interesting products for users. Experimental results show that the hybrid recommendation algorithm can provide a better service experience for users [6]. Jiao W et al. faced the problems of data sparsity and cold start in traditional recommendation algorithms and used K-means, weighting for optimization, while introducing adjustment factors to combine collaborative filtering with dichotomous networks. Relevant experimental data show that the hybrid recommendation algorithm studied is operable and has better recommendation accuracy than the comparison algorithm [7]. Liu et al. propose a personalized service recommendation system in order to improve the competitiveness of manufacturing service platforms, and use a hybrid algorithm to solve the composite service problem. In this algorithm, customer preferences are quantified by clustering, and composite services are optimally ranked by ranking genetics. The analysis results of real cases show that the performance of

the studied recommendation algorithm is good and has some practical value [8].

Nafis et al. studied a travel platform based on a hybrid recommendation algorithm in order to help tourists make personalized travel plans. The platform can recommend relevant resources according to tourists' preferred tourist attractions, and through relevant experimental analysis, the platform provides high accuracy of information resources, which can effectively promote the development of read tourism [9]. Pirasteh et al. team members addressed the situation that collaborative filtering has a cold start, which leads to poor recommendation effect, by capturing various similarities between items and finding out hidden preferences in items to alleviate the problem due to the number of items. The team members have been able to mitigate the low quality of recommendations due to insufficient number of items by capturing various similarities between items and finding hidden preferences in items. Simulation experimental data show that the diversity of similarities can provide more reliable results for users and achieve high quality personalized recommendations [10]. Hu The researchers studied an improved particle swarm optimization algorithm based on multiple criteria combined with diverse adaptations, while using bacterial foraging to improve the convergence of the algorithm in the face of problems such as multi-objective contradictions and unsatisfied constraints in solving hybrid recommendation models. The experimental results show that the convergence and diversity of the studied algorithm are better than the comparison algorithm, and the results of the recommended model solving are highly accurate, have wide coverage, and can provide users with diverse results [11]. Ajaegbu scholars propose an algorithm to improve similarity measurement based on traditional measures in order to alleviate the problems of data sparsity and cold start of collaborative filtering. The performance of the algorithm is analyzed on different data, and experimental results show that the algorithm retains the advantages of existing measures while mitigating the disadvantages of traditional methods [12].

By summarizing the achievements of domestic and foreign researchers, researchers have proposed different improvement measures for the shortcomings of existing recommendation algorithms. Among them, the hybrid recommendation algorithm can effectively integrate the advantages and disadvantages of different recommendation algorithms to achieve accurate and efficient target recommendation, but the contradiction between accuracy and diversity needs further research. Therefore, the research will combine content-based recommendation and collaborative filtering to form a cascaded hybrid recommendation algorithm in order to achieve highly accurate recommendation results. At the same time, in the face of the contradiction between diversity and accuracy, the immune optimization algorithm is introduced into the recommendation results. It is expected that the multi-objective hybrid recommendation algorithm studied can achieve accurate recommendation and multiple needs of users.

### III. HYBRID RECOMMENDATION ALGORITHM FOR CLOUD DATA BASED ON MULTI-OBJECTIVE COLLABORATIVE FILTERING

#### A Recommendation Algorithms based on Own Preference Models

The rapid development of information technology has intensified the problem of data overload, and the massive amount of data has caused many problems for users. In the constant search for solutions, personalized recommendation has emerged, and recommendation systems can effectively filter distracting information and improve data usage [13-14]. Collaborative filtering is a classical algorithm in recommendation systems. The core idea is to find the nearest neighbors by calculating the similarity, and to find the content of interest to the user in the nearest neighbors and recommend it to the target user. Commonly used calculation formulas are such as modified cosine similarity, Pearson, Jaccard similarity, etc. Through similar user groups, some potentially preferred content will also be recommended to the target customer, reflecting the diversity of recommendation results of collaborative filtering [15-16]. As the data expands, the likelihood of different users being interested in the same content among themselves gradually decreases, and the problem of data sparsity cannot be avoided. In addition, when a new user or item appears, relevant attribute information does not exist and the system is unable to make a recommendation resulting in a cold start problem. In the field of textual data recommendation, Content Based Recommendations (CBR) is based on the analysis of users' historical data, and by constructing interest models, the similarity between the interest model and the content of the item reflects the user's preference for the content. However, the recommendation results lack some diversity and cannot effectively explore the potential preferences of users [17]. Therefore, in the field of text data recommendation, research will combine CBR and CF to form a hybrid recommendation algorithm, effectively integrate the advantages of the two, and achieve personalized and diversified content recommendation. At the same time, there is a contradiction between the diversity and accuracy of the recommendation results. In order to alleviate the contradiction between the two, the text data recommendation problem is transformed into a multi-objective optimization problem through modeling, and the immune optimization algorithm is used to optimize the candidate text set to generate the final recommendation list.

The Self Preference Model (SPM) is built on the basis of CBR, which can intuitively reflect the user's own interests, and the whole construction process is shown in Fig. 1. The text features are obtained by pre-processing the user's browsing history, using a vector space model to represent the text features, introducing a time factor to adjust the text model, and finally obtaining the user's own preference model.

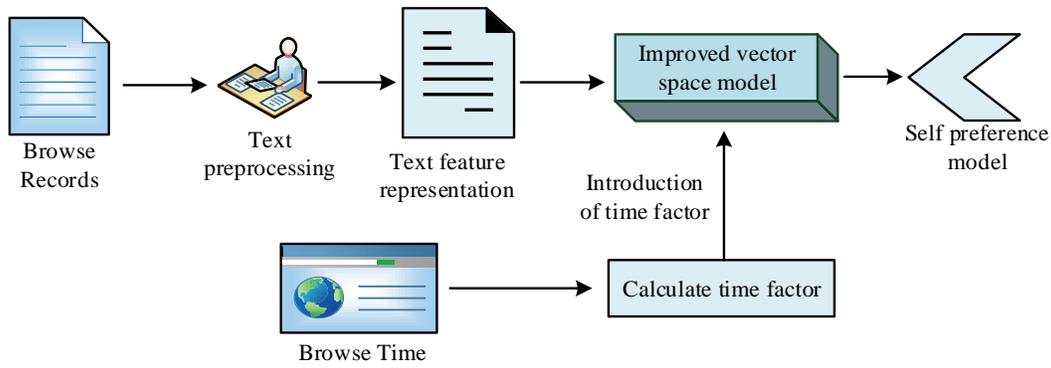


Fig. 1. Construction process of self-preference model.

In the text pre-processing stage, special characters such as emoticons and face characters are removed, and then the text is divided into sequences of word order using the word separation technique jieba. To avoid excessive word sequences that increase the computational effort, meaningless word sequences such as conjunctions, auxiliaries and stop words are usually screened out of the feature word sequences, while verbs, nouns and adjectives are retained in the word sequences. Representation of text features is a key technique in CBR, where key content features are used to represent text information and build a text model. Suppose there is a text dataset  $H = \{h_1, \dots, h_a, \dots, h_m\}$ ,  $m$  is the total amount of text, after the textual sub-word processing, the main feature word order of the text set is  $g$  dimensional composition vector  $F = \{f_1, \dots, f_j, \dots, f_g\}$ , the feature vector of any text  $h_a$  is  $h_a = \{w_{a1}, \dots, w_{aj}, \dots, w_{ag}\}$ ,  $w_{aj}$  represents the weight of the feature word in the text  $h_a$ , the higher the weight value, the more important the feature word  $f_j$  in the text  $h_a$ . The TF-IDF (Term Frequency-Inverse Document Frequency) representation is often used to calculate the weights of feature word sequences in different texts, and the TF is used to count the frequency of feature words in a text set.

$$TF(a, j) = \frac{f(a, j)}{\max Others(a, j)} \quad (1)$$

In eq. (1),  $TF(a, j)$  is the frequency of the feature word in the text  $h_a$ ,  $f(a, j)$  is the absolute frequency of the feature word in the text  $h_a$ , and  $\max Others(a, j)$  is the most frequent feature word in the text  $h_a$ . The IDF is used to represent the distribution of the feature word in the document and to count the frequency of the feature word, which is calculated in eq. (2).

$$IDF(j) = \log\left(\frac{M}{M(j)}\right) \quad (2)$$

In eq. (2),  $M$  represents the number of all texts and  $M(j)$  represents the number of texts where the feature term  $j$  is present.  $IDF(j)$  The larger the number, the lower the number of occurrences of the feature word in the text set. The formula for calculating the weight of a feature word  $w_{aj}$  is shown in eq. (3).

$$w_{aj} = TF(a, j) * IDF(j) \quad (3)$$

After calculating the weights of all feature words, the weight matrix  $HM$  of the text set  $H$  is obtained as shown in eq. (4).

$$HM = \begin{bmatrix} w_{11} & w_{12} & \dots & w_{1g} \\ w_{21} & w_{22} & \dots & w_{2g} \\ \vdots & \vdots & \dots & \vdots \\ w_{m1} & w_{m1} & \dots & w_{mg} \end{bmatrix} \quad (4)$$

Using the text that the user has read as a feature vector, the weight matrix  $HM_u$  for the set of texts that the user  $u$  has read is shown in eq. (5).

$$HM_u = \begin{bmatrix} w_{u11} & w_{u12} & \dots & w_{u1g} \\ w_{u21} & w_{u22} & \dots & w_{u2g} \\ \vdots & \vdots & \dots & \vdots \\ w_{um1} & w_{um1} & \dots & w_{umg} \end{bmatrix} \quad (5)$$

As users' interests change, text recommendations need to be time-sensitive, and analyzing users' recent behavioral data can improve prediction accuracy. Therefore, the study introduces a time-factor adjustment model matrix to obtain users' real-time interest preferences. Define the set of texts read by the user  $u$   $H_u$ ,  $t_0$  is the current time,  $t_i$  is the user's browsing time, and the time factor is calculated by eq. (6).

$$\varepsilon_i = \frac{e^{-(t_0-t_i)}}{\sum_{h_{uj} \in H_u} e^{-(t_0-t_j)}} \quad (6)$$

In eq. (6)  $e$  is a constant. The introduction of a time factor to adjust the weights of the weight matrix  $HM_u$  gives the user's own preference model as  $SPM_u = (W1_{u1}, W1_{u2}, \dots, W1_{ug})$ , where  $W1_{ua} = \sum_{h_{aa} \in H_u, w_{uaj} \in HM_u} \epsilon_i * w_{uaj}$ .

**B Hybrid Recommendation Algorithm with Fused Preference Models**

In practice, if only one's own preferences are considered, the recommendation results lack a certain degree of richness, and it is necessary to explore the potential preferences of users to improve user satisfaction [18]. The Potential Preference Model (PPM) is built by extracting the user's browsing history to form a matrix of user behavior and finding nearest neighbors through a mixture of behavioral and content similarity calculations to solve the problem of not being able to categories similar users due to text diversity [19]. When using CF for recommendation, it is transformed to recommend the feature words of interest to the nearest neighbors, which can effectively avoid the cold start problem, and the construction process is shown in Fig. 2.

In the user behavior matrix, suppose the user set is  $U = \{u_1, u_2, \dots, u_n\}$ ,  $n$  is the total number of users and the text set is  $H = \{h_1, h_2, \dots, h_m\}$ , according to the user's browsing history, when the user  $u$  reads the text  $a$ , then  $x_{ua} = 1$ , and vice versa, the user's behaviour matrix  $X_{n \times m}$  can be expressed as eq. (7).

$$X_{n \times m} = \begin{bmatrix} x_{11} & \dots & x_{1a} & \dots & x_{1m} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{u1} & \dots & x_{ua} & \dots & x_{um} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{n1} & \dots & x_{na} & \dots & x_{nm} \end{bmatrix} \quad (7)$$

The user's level of interest will change over time. To calculate the similarity of user behavior if only the same behavior exists between users is considered, ignoring the time

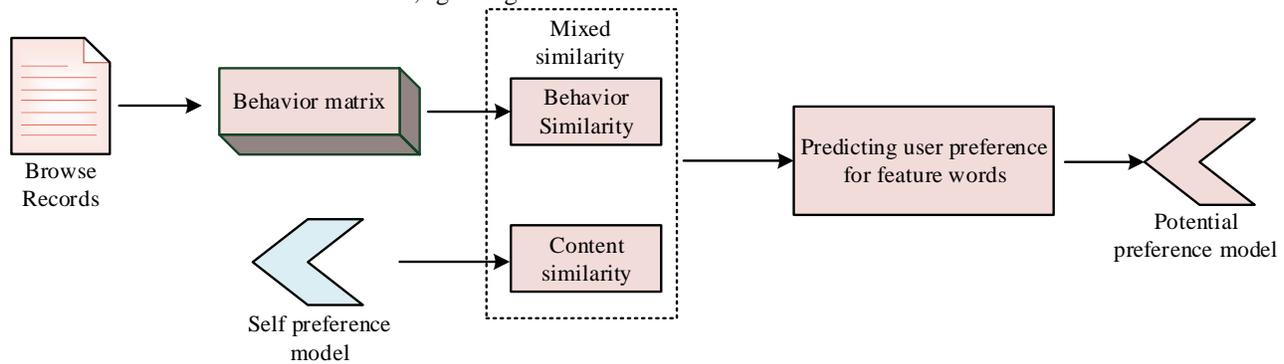


Fig. 2. Potential preference model construction process.

difference in user browsing will affect the level of interest in text recommendations, the time difference in user browsing behavior is defined as time decay, and its formula is eq. (8).

$$t_a = 1 / (1 + \delta |t_{ua} - t_{ra}|) \quad (8)$$

In eq. (8),  $\delta$  represents the time decay factor,  $t_{ua}$  and  $t_{ra}$  represent the time that users  $u$  and  $r$  spent viewing the same text  $a$ , and  $|t_{ua} - t_{ra}|$  represents the time difference between users  $u$  and  $r$  viewing the text  $a$ . The smaller the difference, the greater the similarity between users. At the same time, the influence of the external environment can also lead to user interest, for example, the coverage of popular events will make users interested in the field for a short period of time, and after the heat has passed, users' interest will return to its original state. Therefore, we need to take the external heat factor into account, and define the heat difference between users reading the same text at different times as  $k_a$ , which is calculated by eq. (9).

$$k_a = 1 / \log(1 + |N(a)|) \quad (9)$$

In eq. (9),  $N(a)$  represents the set of users who have viewed the text  $a$ . The more people read the text, the hotter the text is and the more likely users are to view the text due to its hotness, and the larger  $N(a)$  is, the greater the difference in hotness  $k_a$  is. Therefore, the formula for calculating the similarity of user behavior, taking into account the temporal differences in the generation of browsing behavior and the heat of the text, is eq. (10).

$$sim_{act}(u, r) = \frac{\sum_{a \in H(u) \cap H(r)} \eta * t_a + (1 - \eta) * k_a}{\sqrt{|H(u)| |H(r)|}} \quad (10)$$

In eq. (10),  $H(u)$  and  $H(r)$  represent the set of texts viewed by the user  $u$  and  $r$ , and  $\eta$  represents the moderator, with larger values indicating a greater influence of temporal differences and smaller values indicating a greater influence of text hotness,  $\eta \in [0,1]$ . Content similarity is calculated by the own preference model, and the user's content similarity is calculated by eq. (11).

$$sim_{con}(u,r) = \frac{PSM_u * PSM_r}{|PSM_u| \times |PSM_r|} \quad (11)$$

By combining the behavioural similarity with the content similarity through a weighting factor, the formula for the mixed similarity is eq. (12).

$$sim(u,r) = \chi sim_{act}(u,r) + (1-\chi) sim_{con}(u,r) \quad (12)$$

In eq. (12),  $\chi$  represents the weighting factor and  $\chi \in [0,1]$ . When  $\chi = 0$  only content features are considered, the larger  $\chi$  is, the greater the degree of influence of user behavioural features. The set of nearest neighbours of the user  $u$  is  $N(u)$  and the potential preference model is  $PPM_u = (W2_{u1}, W2_{u2}, \dots, W2_{ug})$ , where  $W2_{ua} = \sum_{r \in N(u)} \frac{sim(u,r)}{\sum_{r \in N(u)} sim(u,r_x)} W1_{ra}$  is used to find the nearest neighbors of the user's mixed similarity.

SPM can reflect the user's own preferences and has obvious personalization characteristics, while PPM can expand the user's interests and has diversity characteristics. The fusion of the two can ensure that the recommended content can meet the user's own preferences and enhance the diversity of content. When fusing SPM and PPM, the measurement between personalization and diversity should not be simply mixed.  $\max()$  function can take the feature vector with greater weight in the two models as the final preference vector of the user, which is selected as the fusion function, and the Fusion Preference Model (FPM) constructed is  $FPM_u = (W3_{u1}, W3_{u2}, \dots, W3_{ug})$ , where  $W3_{ua} = \max(W1_{ua}, W2_{ua})$ .

A fusion preference model is used to calculate the similarity between users, select users with greater similarity as nearest neighbors, predict the interest of users in candidate text sets from the browsing records of nearest neighbors, and recommend results according to the degree of preference, and the preference prediction formula is eq. (13).

$$P(u,a) = \sum_{r \in N(u)} sim(u,r) x_{ra} \quad (13)$$

In eq. (13),  $x_{ra}$  indicates whether the user  $r$  has generated a view record for the text  $a$ , and if so,  $x_{ra} = 1$ , and if not,  $x_{ra} = 0$ .

### C NNIA-based Multi-objective Optimization Recommendation Algorithm

In recommendation systems, there is a conflicting status quo between the diversity and accuracy of recommendation results. To alleviate the conflict between the two, the text data recommendation problem is transformed into a multi-objective optimization problem through modelling. Multi-objective optimization problems are common in everyday life. In solving multi-objective problems, the objectives are so interlocked that finding the best solution that satisfies all the objectives is impossible, so a trade-off between the objectives is needed to find the relatively better solution [20]. In the field of textual data recommendation, the first consideration is the accuracy of the recommendation results. High accuracy can improve user satisfaction and also measure the level of interest in the recommendation results. The objective function for accuracy is chosen as the similarity matching function and is calculated by eq. (14).

$$\begin{cases} y_v = \frac{1}{L} \sum_{c \in R} 1 - sim(c, P_u) \\ sim(c, P_u) = \frac{1}{|P_u|} \sum_{a \in P_u} sim(c, a) \end{cases} \quad (14)$$

In eq. (14),  $L$  represents the length of the list of all recommendations,  $P_u$  represents the set of texts that the user  $u$  has viewed, and  $sim(c, a)$  represents the similarity between the text  $c$  and the text  $a$ .  $y_v$  The smaller the value, the more similar the recommendation is to the text the user has viewed and the better the accuracy of the recommendation. To help users find more interesting text data, diversity is taken into account and the objective function of diversity is eq. (15).

$$y_D = \frac{1}{L(1-L)} \sum_{c,a \in R, c \neq a} sim(c, a) \quad (15)$$

From eq. (15), the smaller  $y_D$  is, the smaller the similarity between the texts in the recommendation list, which means the richness and diversity of the recommendation content is better. Considering diversity and accuracy, it is found that the smaller the  $y_v$  and  $y_D$ , the better. Therefore, the final objective function is minimization, and the specific expression is eq. (16).

$$Min : Y = \{y_v, y_D\} \quad (16)$$

The Non-dominated Neighborhood Immune Algorithm (NNIA) has excellent evolutionary performance and is used to solve multi-objective problems due to its fast convergence and robustness. A certain proportion of each superior antibody is cloned, and the population is then crossed and mutated to enhance the diversity of the population and avoid local optima. During the crossover operation, the same gene loci in the two parent individuals are inherited, and the remaining positions are randomly crossed with a probability of [0,1]. The mutation operation is carried out for the parent individuals. The crossover and mutation operations are shown in Fig. 3.

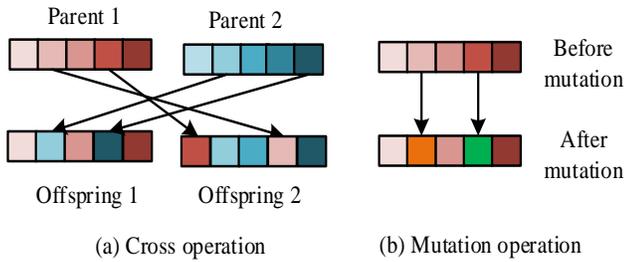


Fig. 3. Crossover and mutation operation.

Using the NNIA algorithm, it is necessary to encode the texts in the candidate set with real numbers. Let the encoding of an antibody be  $z\{z_1, K, z_b, K, z_l\}$ ,  $z$  denotes any user,  $z_b$  denotes the recommended text number, and each text has a unique number. The candidate text set obtained from the fusion preference model is optimized by crossover and variation operations to produce the final recommendation list. The

specific process of text data recommendation using NNIA algorithm is shown in Fig. 4. The user text candidate set data program generated by the fusion preference model is used to encode the candidate text set in real numbers. Set the maximum number of iterations, population size, crossover and mutation probability and other related parameters to randomly generate the initial antibody population. Identify the dominant antibody and clone the dominant antibody. Place the cloned individuals in the temporary population. Select the individuals with the lowest density according to the population density to enter the new population and update the dominant population. The new dominant population is cloned, crossed, and mutated to continue to update the dominant population until the maximum number of iterations is reached, and finally the best recommendation list of text recommendation is output.

The NNIA algorithm can be used to generate a series of recommendation lists with different recommendation targets, personalized to the specific needs of the user, increasing user satisfaction and loyalty.

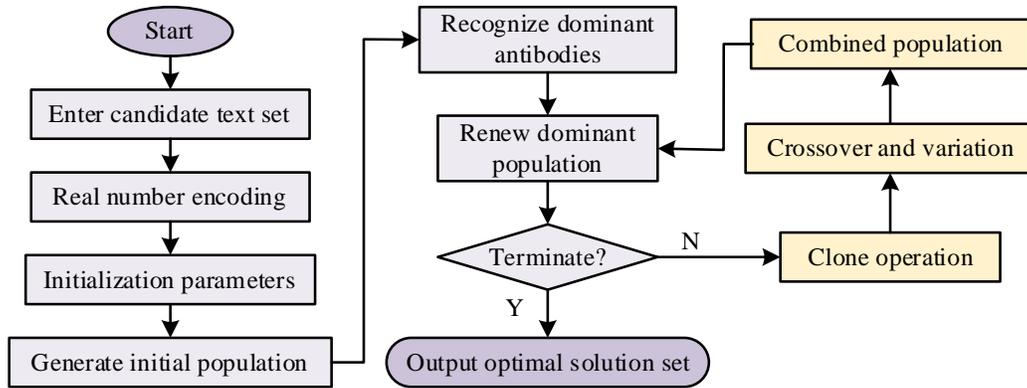


Fig. 4. NNIA algorithm text recommendation process.

#### IV. PERFORMANCE ANALYSIS OF MULTI-OBJECTIVE COLLABORATIVE FILTERING-BASED DATA-MIXING RECOMMENDATION ALGORITHMS

##### A Performance Validation of Hybrid Recommendation Algorithms Incorporating Preference Models

In order to determine the values of the parameters present in the latent preference model (PPM), namely the temporal

decay factor  $\delta$ , the text heat modifier  $\eta$  and the mixed similarity weighting factor  $\chi$ , Fig. 5 shows the variation in the accuracy of the recommendation results for different temporal decay factors  $\delta$  and modifiers.  $\eta$

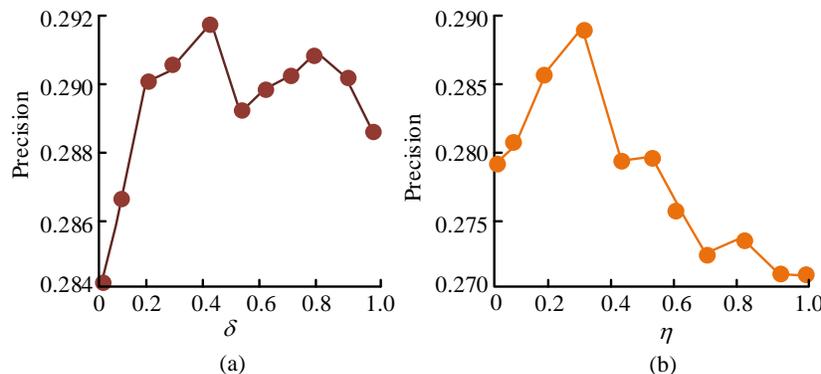


Fig. 5. Recommended accuracy when  $\delta$  and  $\eta$  have different values.

As can be seen in Fig. 5(a), the accuracy of the recommendations fluctuates as the value of  $\delta$  increases, and the experimental results show that there is a maximum accuracy of the recommendations at  $\delta=0.4$ . In Fig. 5(b), as the value of  $\eta$  increases, the recommended accuracy curve increases to a peak and then gradually decreases, showing the maximum accuracy at  $\eta=0.3$ . In the case of  $\delta=0.4$  and  $\eta=0.3$ , the best weighting factor  $\chi$  was analysed, and the accuracy curves for different values of  $\chi$  are shown in Fig. 6.

As can be seen in Fig. 6, the best accuracy of the recommended results is obtained at  $\chi=0.7$ . The experimental analysis finally resulted in the accurate values of each parameter when constructing the PPM, i.e.  $\delta=0.4$ ,  $\eta=0.3$ ,  $\chi=0.7$ . In order to investigate whether the recommendation performance of the fusion preference model (CBCF) is better than before fusion, the study conducted comparison experiments with SPM and PPM, and Fig. 7 shows the comparison results of the accuracy and recall of the three models' recommendations.

As can be seen in Fig. 7(a), the trend of all three models decreases as the amount of recommended text increases, but the accuracy of CBCF is always higher than the comparison algorithms, with a maximum increase of 9.57% and 8.23% over the accuracy of SPM and PPM. As can be seen in Fig. 7(b), the pattern of change in recall is opposite to the pattern of change in accuracy, with the higher the number of recommended texts, the greater the recall. the recall of CBCF is higher than the other two models, with a 9.97% and 7.65% improvement over the SPM and PPM recall. The fusion of

models can effectively improve the accuracy of recommendations. Fig. 8 shows the F value and recommended coverage results of the three models.

As can be seen from Fig. 8(a), the change trend of the F value curve of the three models is similar. With the increase of the recommended text, the F value rises to the highest value, and then decreases gradually. The maximum F value of CBCF is 0.493, and the maximum values of SPM and PPM are 0.438 and 0.461 respectively. Compared with the comparison model, the F value of CBCF is 0.055 and 0.032 higher. Fig. 8(b) shows the results of recommendation coverage. It can be seen that with the increase of the number of recommended texts, the coverage of the three models is on the rise, and the coverage of CBCF is always higher than that of the comparison model. Compared with SPM and PPM, the coverage increased by 10.04% and 16.45%. Relevant data shows that the fusion of models can effectively improve recommendation performance.

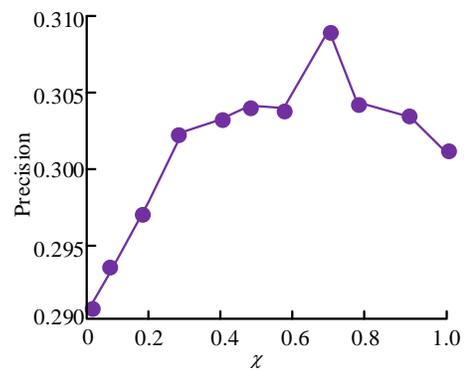


Fig. 6.  $\chi$  Recommended accuracy rate for different values.

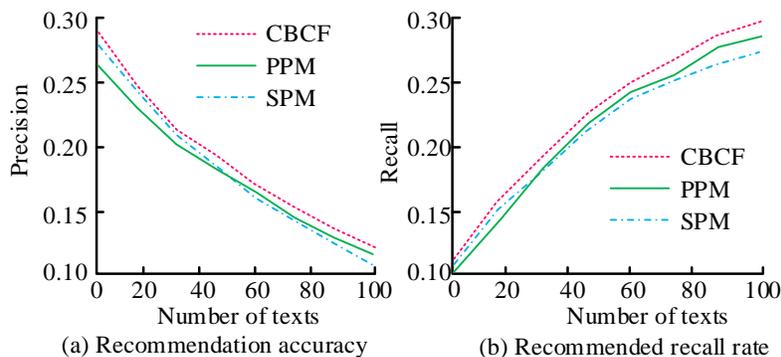


Fig. 7. Comparison results of accuracy and recall rate recommended by three models.

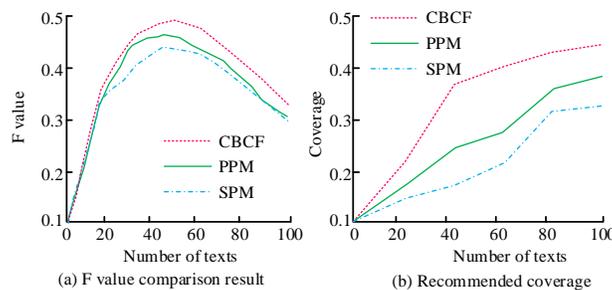


Fig. 8. F value and recommended coverage results of the three models.

**B Analysis of the Effectiveness of the Hybrid Recommendation Algorithm based on NNIA Multi-objective Optimization**

To verify the performance of the NNIA multi-objective optimization recommendation algorithm, the maximum number of iterations of the algorithm was set to 300, the number of nearest neighbors to 25 and the size of the dominant population to 50. The top 50 non-dominated solutions were selected as the recommendation result solution in NNIA according to the degree of density, and one user was randomly selected for presentation, and the results are shown in Fig. 9.

The distribution of the user 27 recommended solution set in the target space is illustrated in Fig. 9, where it can be seen that the NNIA optimized recommendation algorithm can effectively obtain 50 different solutions. The leftmost solution is the one with the highest accuracy and the worst diversity among all solutions, while the rightmost solution is the one with the best diversity and the lowest accuracy among all solutions. From the distribution of the solution set, as the value of the diversity function increases, the value of the accuracy function decreases, indicating that there is a contradiction between the two indicators in the recommendation process, which is in line with the law of the recommendation process and verifies the effectiveness of the NNIA multi-objective algorithm.

To analyses the performance of the studied NNIA multi-objective optimized hybrid recommendation algorithm (CBCF-NNIA), it was compared with Recommendations Based on User Browsing (BUB), Personalized recommendation based on user interest (PBUI) and Hybrid recommendation with fused preference model (CBCF) in terms of accuracy, diversity of recommendation results. The accuracy and diversity of recommendation results are compared. The accuracy and diversity of recommendation results are compared.

As can be seen in Table I, the mean accuracy of the recommendation lists of 9 out of 10 users outperformed the other three comparison algorithms, and the maximum accuracy of the recommendation lists of the remaining 1 user outperformed the comparison algorithms. The experimental

data shows that the recommendation list given by the CBCF-NNIA algorithm studied is not necessarily the most accurate, but can give more accurate recommendations than the other three comparison algorithms.

Diversity can indicate the variability between texts in a list. The smaller the diversity value, the greater the similarity between texts and the richer the content. Table II shows the recommended diversity results for several algorithms.

As can be seen in Table II, the mean value of the diversity of the recommendation lists of 8 out of 10 users is lower than the other three comparison algorithms, and the minimum value of the diversity of the recommendation lists of the remaining 2 users is lower than the comparison algorithms. The experimental data suggests that the CBCF-NNIA algorithm does not give the best diversity of recommendation lists, but can give richer recommendations than the other three comparative algorithms. The study selected 50 sets of experimental data to calculate the average accuracy and average diversity of several recommendation algorithms, and the overall comparison results are shown in Table III.

As can be seen in Table III, the accuracy and diversity of the recommended results of the studied CBCF-NNIA algorithm outperform the comparison algorithm, which provides users with textual content that meets their needs as well as being rich and diverse.

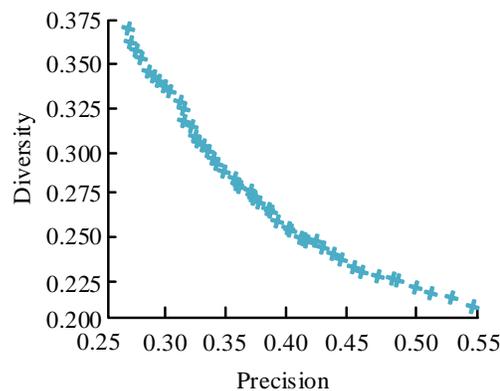


Fig. 9. Non dominated solution of user 27.

TABLE I. RECOMMENDATION ACCURACY OF SEVERAL RECOMMENDATION ALGORITHMS

User ID	CBCF-NNIA			CBCF	BUB	PBUI
	min	max	mean			
300	0.2	0.4	0.32	0.3	0.2	0.3
301	0.2	0.5	0.39	0.3	0.2	0.2
302	0.1	0.4	0.26	0.2	0.1	0.2
303	0.3	0.7	0.45	0.6	0.3	0.5
304	0.2	0.4	0.27	0.2	0.2	0.2
305	0.2	0.6	0.36	0.3	0.2	0.3
306	0.3	0.7	0.51	0.4	0.3	0.4
307	0	0.5	0.23	0.2	0.1	0.2
308	0.5	0.9	0.73	0.7	0.6	0.7
309	0	0.2	0.14	0.1	0	0.1

TABLE II. RECOMMENDATION DIVERSITY RESULTS OF SEVERAL ALGORITHMS

User ID	CBCF-NNIA			CBCF	BUB	PUBI
	min	max	mean			
300	0.36	0.45	0.39	0.40	0.42	0.43
301	0.32	0.42	0.36	0.36	0.37	0.40
302	0.24	0.29	0.26	0.28	0.30	0.39
303	0.30	0.39	0.36	0.38	0.38	0.41
304	0.25	0.33	0.28	0.31	0.29	0.38
305	0.29	0.33	0.31	0.31	0.32	0.35
306	0.26	0.31	0.29	0.32	0.34	0.37
307	0.28	0.38	0.33	0.40	0.40	0.45
308	0.27	0.32	0.29	0.30	0.33	0.31
309	0.31	0.37	0.35	0.36	0.39	0.47

TABLE III. OVERALL COMPARISON RESULTS OF SEVERAL RECOMMENDED ALGORITHMS

/	CBCF-NNIA	CBCF	BUB	PUBI
Average accuracy	0.35	0.32	0.29	0.30
Average diversity	0.29	0.31	0.36	0.34

## V. DISCUSSION AND CONCLUSION

The research determines the parameter values in the PPM model through the accuracy changes of the recommendation results under different parameter values. The results show that the recommended accuracy of PPM is the best when  $\delta = 0.4$ ,  $\eta = 0.3$ ,  $\chi = 0.7$ . The PPM considers the impact of user browsing time difference, external environment and other user behavior characteristics on user interest, and the recommendation results will be more accurate. By analyzing the accuracy, recall, F value and coverage of the recommendation results, the recommendation performance of the three models is compared. The results show that the various index values of CBCF are higher than those of the comparison model. This is because CBCF is a model integrating SPM and PPM models, which can ensure that the recommended content not only meets the user's own preferences, but also improves the content diversity. When the model is fused, the personalization and diversification are measured, and the feature vector with larger weight in the model is used as the user's final preference vector, so as to construct the CBCF model. Therefore, CBCF model combines the advantages of the two models and can effectively improve its recommendation performance. The performance of CBCF-NNIA algorithm is analyzed from the diversity and accuracy of recommendation results. The results show that the algorithm can better take into account the diversity and accuracy of recommendation results, and meet the user's personalized and rich and diverse needs of text. In order to solve the contradiction between recommendation diversity and accuracy, the problem is transformed into a multi-objective optimization problem. The NNIA algorithm has fast convergence speed and good robustness. It uses crossover and mutation operations to enhance the diversity of the population and avoid the occurrence of local optimization. Introducing

NNIA into CBCF can effectively obtain the content list with high accuracy and good diversity.

The rapid development of Internet technology has made the problem of information overload increasingly serious, and recommendation systems are one of the effective techniques to improve information utilization. Faced with the shortcomings in existing recommendation algorithms and the conflict between recommendation accuracy and diversity, the study combines recommendation algorithms with multi-objective optimization algorithms and proposes an NNIA multi-objective optimization hybrid recommendation algorithm. Experimental tests determine the relevant parameters of the hybrid recommendation model. The results of the comparison experiments showed that the accuracy and recall of the hybrid recommendation algorithm were higher than those of the model before the hybridization, with the accuracy improving by 9.57% and 8.23% over SPM and PPM respectively, and the recall improving by 9.97% and 7.65% respectively. The results are consistent with the personalized recommendation law. By exploring the performance of CBCF-NNIA recommendation accuracy and diversity, the results show that the algorithm can effectively obtain content lists with high accuracy and good diversity, which can meet the multiple needs of different users and improve user satisfaction. Although the research has achieved certain results, there are still many shortcomings. The recommendation process mainly considers textual content data, and subsequently, factors such as text labels and categories are also taken into account, while the problem of data sparsity has not been effectively solved, which will be investigated in the future.

## REFERENCES

- [1] H. Cao, "Personalized web service recommendation method based on hybrid social network and multi-objective immune optimization." *Journal of Information Processing Systems*, vol. 17(2), pp. 426-439, 2021.
- [2] H. Pan, Z. Zhang, "Research on context-awareness mobile tourism e-commerce personalized recommendation model". *Journal of Signal Processing Systems*, vol. 93(2), pp. 147-154, 2021.
- [3] S. Li, R. Chen, C. Sun, H. Yao, X. Cheng, Z. Li, X. Kang, "Region-aware neural graph collaborative filtering for personalized recommendation." *International Journal of Digital Earth*, vol. 15(1), pp. 1446-1462, 2022.
- [4] Y. Liu, X. You, S. Liu, "Multi-ant colony optimization algorithm based

- on hybrid recommendation mechanism.” *Applied Intelligence*, vol. 52(8), pp. 8386-8411, 2022.
- [5] Y. Tian, B. Zheng, Y. Wang, Y. Zhang, Q. Wu. “College Library Personalized Recommendation System Based on Hybrid Recommendation Algorithm.” *Procedia Cirp*, vol. 83(4), pp. 490-494, 2019.
- [6] Y. Wang, Y. Zhu, Z. Zhang, H. Liu, P. Guo. “Design of Hybrid Recommendation Algorithm in Online Shopping System.” *Journal of New Media*, vol. 3(4), pp. 119-128, 2021.
- [7] W. Jiao, B. Ma, L. Zhu “Hybrid Recommendation Algorithm Based on Improved Collaborative Filtering and Bipartite Network”, *ICBDT 2020: 2020 3rd International Conference on Big Data Technologies*. vol. 11(4), pp. 80-84, 2020.
- [8] Z. Liu, L. Wang, X. Li, S. Pang. “A multi-attribute personalized recommendation method for manufacturing service composition with combining collaborative filtering and genetic algorithm.” *Journal of Manufacturing Systems*, vol. 58(A), pp. 348-364, 2021.
- [9] F. Nafis, K. Al Fararni, B. Aghoutane, A. Yahyaouy, J. Riffi, A. Sabri. “Hybrid recommender system for tourism based on big data and AI: A conceptual framework.” *Big Data Mining and Analytics*, vol. 4(1), pp. 47-55. 2021.
- [10] P. Pirasteh, M. R. Bouguelia, K. C. Santosh, “Personalized recommendation: an enhanced hybrid collaborative filtering.” *Advances in Computational Intelligence*, vol. 1(4), pp. 1-8, 2021.
- [11] Z. Hu, Y. Lan, Z. Zhang, X. Cai, “A many-objective particle swarm optimization algorithm based on multiple criteria for hybrid recommendation system.” *KSII Transactions on Internet and Information Systems (TIIS)*, vol. 15(2), pp. 442-460, 2021.
- [12] C. Ajaegbu, “An optimized item-based collaborative filtering algorithm.” *Journal of ambient intelligence and humanized computing*, vol. 12(12), pp. 10629-10636, 2021.
- [13] Na L, Ming-xia L, Hai-yang Q, Hao-long S. A hybrid user-based collaborative filtering algorithm with topic model. *Applied Intelligence*, 2021, 51(11): 7946-7959.
- [14] Aljunid M F, Huchaiah M D. An efficient hybrid recommendation model based on collaborative filtering recommender systems. *CAAI Transactions on Intelligence Technology*, 2021, 6(4): 480-492.
- [15] Y. Xin, B. Henan, J. Niu, W. Yu, H. Zhou, X. Ji, P. Ye, “Coating matching recommendation based on improved fuzzy comprehensive evaluation and collaborative filtering algorithm.” *Scientific Reports*, vol. 11(1), pp. 1-14, 2021.
- [16] L. Xie, Z. Hu, X. Cai, W. Zhang, J. Chen, “Explainable recommendation based on knowledge graph and multi-objective optimization.” *Complex & Intelligent Systems*, vol. 7(3), pp. 1241-1252, 2021.
- [17] F. Rezaimehr, C. Dadkhah “A survey of attack detection approaches in collaborative filtering recommender systems.” *Artificial Intelligence Review*, vol. 54(3), pp. 2011-2066, 2021.
- [18] L. Duan, W. Wang, B. Han, “A hybrid recommendation system based on fuzzy c-means clustering and supervised learning.” *KSII Transactions on Internet and Information Systems (TIIS)*, vol. 15(7), pp. 2399-2413, 2021.
- [19] Thakker U, Patel R, Shah M. “A comprehensive analysis on movie recommendation system employing collaborative filtering.” *Multimedia Tools and Applications*, vol. 80(19), pp. 28647-28672, 2021.
- [20] J. Chen, B. Wang, Z. Ouyang, Z. Wang, “Dynamic clustering collaborative filtering recommendation algorithm based on double-layer network.” *International Journal of Machine Learning and Cybernetics*, 2021, vol. 12(4), pp. 1097-1113, 2021.

# Equally Spread Current Execution Load Modelling with Optimize Response Time Brokerage Policy for Cloud Computing

Anisah Hamimi Zamri<sup>1</sup>, Nor Syazwani Mohd Pakhrudin<sup>2</sup>, Shuria Saaidin<sup>3</sup>, Murizah Kassim<sup>4\*</sup>

School of Mechanical Engineering-College of Engineering, Universiti Teknologi MARA, Shah Alam, Selangor, Malaysia<sup>1</sup>  
School of Electrical Engineering-College of Engineering, Universiti Teknologi MARA, Shah Alam, Selangor, Malaysia<sup>2,3,4</sup>  
Institute for Big Data Analytics and Artificial Intelligence (IBDAAI), Universiti Teknologi MARA, Shah Alam, Selangor, Malaysia<sup>4</sup>

**Abstract**—Cloud computing is one of the significant technologies that is used to provide seamless internet surfing for large-scale applications and data storing purposes. The cloud is described as a large platform that enables users to access data from the internet without needing to buy storing space in their equipment such as a computer. Many studies have analysed the load-balancing technique on the cloud to distribute tasks equally between servers using the Equally Spread Current Execution (ESCE) algorithm. ESCE, which is a dynamic load balancer, has quite a few problems such as average level performance and too long of response time which affected the Quality of Services. This research has simulated a cloud computing concept using the ESCE Load Modelling technique with the CloudAnalyst simulator for three servers of Data Center (DC) locations. The ESCE was simulated to enhance its algorithm's performance as a load balancer and higher throughput in the cloud environment. The result shows that ESCE average overall response time is shortest when the DC is located at R0 with response times of 15.05s, 13.05s with 10 VMs, and 8.631s with the Optimize Response Time brokerage policy. This research is significant to promote notable load-balancing technique testing for virtualized cloud machines data centers on Quality of Services (QoS) aware tasks for Internet of Things (IoT) services.

**Keywords**—Equally spread current execution (ESCE); optimize response time brokerage; cloud computing; load balancer; data modelling

## I. INTRODUCTION

Pay-as-you-go online computing services, such as apps, storage, and processing power, are referred to as cloud computing. Due to how simple and affordable it is to use large-scale applications, the demand for this technology has greatly increased [1]. A significant aspect of this environment is the scalable delivery of IT infrastructure and applications as a service, according to the definition of cloud computing, which is end user-focused on how they may experience the cloud environment such as data storage for the information technology for education, remote monitoring and mobile robot for surveillance application [2-5].

Fig. 1 presents the illustration of cloud computing as a user device accessing the internet and communication exists between servers, applications, and databases. Cloud computing may be utilised as an interface for the integration to quickly

complete a variety of client requirements, in just one click. The biggest issue with this system is managing a large number of client requests that could total millions at one time[6, 7]. Mainly maintaining and enhancing prior software investments while responding to environmental changes is the major objective of current application adaptation[8]. The private cloud, the public cloud, the hybrid cloud, and the community cloud are the four cloud computing deployment methods that are based on research and are grouped by their distribution and physical location[9]. Three types of cloud service models are Software as a Service (SaaS), Platform as a Service (PaaS), and Infrastructure as a Service (IaaS).

Concerning effectively managing user requests, load-balancing methods must be used [10]. Load-balancing serves as a tool for assignation and also controls the system's overall rendering [11]. The Equally Spread Current Execution (ESCE) is one of the algorithms that act as a load balancer to process user requests. ESCE algorithm (LBA) allocates user requests equally to all virtual machines (VM) bound to the data centre [12]. This kind of load balancer keeps track of the inventory of virtual machines and their current availability. However, depending on the suggested work or goal of the cloud environment, the algorithms' efficacy is unclear [13]. Each algorithm varies depending on the type of load-balancing within the bounds of cloud computing [14]. The turn-based Round Robin load-balancing solution has its unique algorithm [15]. The first server will receive the first request, the second server will receive the second request, and so on. Following distributing tasks after determining their size, the Equally Spread Current Execution algorithm uses a spread spectrum technique.

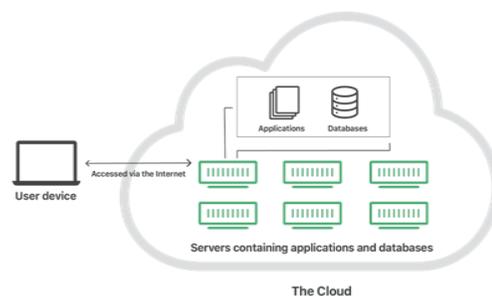


Fig. 1. Illustration of cloud computing

Additionally, this algorithm performs average in terms of response rate and has excessively long mediocre minimum and maximum response times [16]. Response time-wise, the Equally Spread Current Execution algorithm is excessively slow. This is a result of the algorithm's lengthy check of the request size. The algorithms must first successfully assess the availability status of each machine before allocating them to the available virtual machines. This causes system delays and reduces the effectiveness of the entire load-balancing strategy. When compared to other load balancers now in use, the performance of this method is in the middle of the pack [17]. Equally Spread Current Execution was not the best load-balancing strategy when compared to the others. An earlier study compared it to other load-balancing methods including Round Robin and Throttled Load-balancing [18]. It does not, however, qualify as the worst technique either. The issue is that it is not the best technique to employ at the moment.

Cloud computing is Internet-based computing that offers a pool of customizable computing resources such as networks, storage, servers, applications, and services without requiring interaction with the service provider and with little administration effort[19]. The RR protocol's moving horizon estimation problem for a class of discrete time-delay systems. To avoid data collisions, communication between the sensor nodes and the remote state estimator is carried out over a shared network, with only one sensor node able to transmit data at any given time. The RR protocol organizes the transmission order of sensor nodes, with the selected node gaining network access modelled as a periodic function. The device model is reformulated into a linear system without delays due to lifting technology. The problem at hand aims to construct a moving horizon estimator that minimizes the estimation error. In matrix inequality, a proper condition is defined to ensure ultimate boundedness [16]. In cloud computing, the cloud infrastructure cannot handle the flow of information independently with the profusion of data, devices, and interactions [7]. Load-balancing distributes all workloads across each node in a shared or mutual system to maximize resource utilization and reduce job response time. The most important aspect of cloud computing is scheduling, which includes workloads and workflow scheduling under the platform as a service model. The infrastructure as service task to VMs scheduling which machine is decided by the scheduler should go on which job or VMs[20]. Evaluating the energy required for communication between the devices participating in this process and the suggested method's appropriateness for handling optimization problems like VM placement is necessary[21]. Load-balancing is a method of reassigning the entire load to the various nodes of a collaborative system to improve resource efficiency and the job's response time while avoiding a situation where specific nodes are overloaded and others are underloaded. A load-balancing algorithm involved in identity ignores the system's previous state or behavior, relying instead on the system's neighboring behavior. This load can be measured in terms of CPU, memory use, sluggishness, or network load[22]. Load-balancing's primary aim is to enhance device efficiency and functionality for today's QoS for IoT services in communication and data transfer. It is used in cloud computing systems to provide successful alternative solutions in the event of a system failure and to ensure the best possible

use of system components. Load-balancing techniques are divided into two main models Static Model and Dynamic Model. First is Static Load-balancing: Static load-balancing occurs in a static environment where the output of algorithms is unaffected by the system's current state. As a result, user expectations do not change over time [23]. Second is Dynamic Load-balancing: The system's state has a significant impact on balancing the efficiency of the algorithms. Since resources are versatile in a dynamic environment, algorithms efficiently perform load-balancing[24].

An increasing number of computation-intensive and delay-sensitive mobile apps keep appearing alongside the evolution of smart Mobile Devices (MD). The task migration issue in the context of cloudlet federation is the primary topic of this research[25]. Considered a cloudlet federation scenario involving three distinct Cloudlet service Providers (CLP) and a remote cloud. Cloudlet federation can efficiently minimise cloudlet deployment and management expenses by pooling resources among CLPs. Task migration faces new obstacles because CLPs vary in their user and resource counts.

Proactive dynamic VM consolidation is used in this research to improve resource utilisation and performance without sacrificing the energy economy[26]. The suggested algorithm creates a fine-grained categorization that takes workload considerations into account by utilising machine learning techniques to create complementary profiles that reduce cross-application interference by strategically colocating HPC and non-HPC workloads. Real HPC workloads were simulated for the study using CloudSim. The outcomes proved that, in terms of the metrics in important areas, the suggested algorithm beats all heuristic methods[26-28].

This research examined the performance of the Equally Spread Current Execution load modelling method for Quality of Service, QoS aware task placement for the Internet of Things (IoT). Since the ESCE load modelling technique requires communication between the data centre and load balancer, the procedure results in overhead [29]. Intended to examine the behaviour of a large distributed system, this project investigated the study of ESCE load modelling using the CloudAnalyst simulator. The user-friendly interface of the CloudAnalyst simulator made it easier for analysts to set up simulations. This research has successfully analysed the performance of ESCE in various situations and proven that it works efficiently by implementing the correct configuration according to the scale of the application.

The paper is divided into five main sections. The Introduction section provides a brief overview of the problem and research question, as well as an introduction to the relevant literature. The theoretical and proposed work section brief the overview of the research gap and the literature review, including the proposed work and some of the mathematical modelling techniques. The Methodology section outlines the research methods used, including the data sources and analysis technique. The Results section presents the findings of the research, while the Discussion section discusses the implications of the research and its implications for future research. Finally, the Conclusion section summarizes the research and provides suggestions for further research.

## II. THEORETICAL AND PROPOSED WORK

Critical analysis of vast amounts of data, including energy production and consumption, is required to develop a secure and sustainable energy system [30]. Due to the high demand for cloud services, much research has been done to improve the performance of the current existing load-balancing. Upon enabling the stability of cloud environments, load-balancing is a key role to ensure no node in the load-balancing is unequally utilized than the other [31]. A study has proposed a genetic algorithm-based task scheduling for load-balancing which demonstrated that the proposed method results exceed the performance of current load-balancing techniques [32]. Proposed algorithms are known to surpass basic static and dynamic load-balancing since their method has attempted to solve problems by creating an alternative to minimize the response time such as designing hybrid algorithms in taking a part of existing algorithms and attaching additional processes in the algorithm.

Another study analyzed the response time of dynamic load-balancing, their results showed that all algorithms used in the simulation, THR, ESCE and Particle Swarm Optimization (PSO) gave the same results when the resource is homogenous. Otherwise, entitled to swarm intelligence ability that the task can be executed more efficiently[33]. Most proposed work has proven that the existence of a variety of cloud computing aspects needs more systematic and advanced algorithms to cater for the complicated environment [34]. According to a study, analysis is done to compare the performance between three existing load-balancing types namely the Round Robin (RR), Equally Spread Current Execution (ESCE) and Throttle (THR) using CloudAnalyst by broker policy grouping and load-balancing which results showed that THR has the best performance compared to the other two as prevents overload efficiently by thresholding the available VMs list [35]. The analysis is done by different data centre policies in the same simulation scenario with nine possible load-balancing approaches and five different workloads which have obtained 45 different results. Their work is quite similar to this analysis as they reviewed the performance of the load-balancing method to compare the results and determine which one has the best performance analytically.

In another comparative study done to evaluate the performance of algorithms, the increasing number of users has been used as a variable as they claimed much previous research has not considered the increasing number of users as a contributor to the improvement of algorithms whereas in the simulation to compare three algorithms, RR, ESCE, and THR they have placed UB user base and data centres in the same region by neglecting the effect of geographical distance [36]. Their results proved that algorithms performance does increase with user numbers as initially in the simulation with 5000 users, no significant difference can be seen between the three algorithms' results. Another simulation has been done to analyze performance by using different service broker policies using four data centres all located in different regions but having the same amount of VMs which is 50 [37]. Simulation is done alternately for three different broker policies to analyze performance based on response time, request processing time and cost the results given THR has the best performance with a

small processing cost compared to the others. While another study has focused on various policies utilized for load-balancing, a simulation was done between all six different user bases that are utilized with four data centres and in these data centres 25 VMs, 50 VMs 75 VMs, and 100 VMs respectively announces to the 4 DCs [38]. The load-balancing algorithms compared are RR, ESCE, THR and First Come First Serve (FCFS) algorithms. The peak hour users set for those 6 user bases vary with each other and concluded that RR has the best integrated performance. Next is a comparative analysis comparing THR, RR, ESCE, FCFS, and Shortest Job First (SJF) [39]. In this analysis, they optimized 6 user bases in all 6 regions respectively with 4 data centres in R0, R4, R2, and R3 containing 15 physical machines consolidated with 10 VMs each. The results showed that FCFS performed the best when it is compared on a data centre processing time basis while ESCE performed best when it comes to the lowest total cost. They showed that a different basis can give different results.

Analysis of the load modelling utilizing Equally Spread Current Execution (ESCE) load-balancing is the goal of this study. The CloudAnalyst simulator, an expanded piece of software created from the CloudSim toolkit, will be used as the approach abandoned for data collecting. A programmer called CloudSim toolkit is used to track internet behavior based on settings made in the simulation configuration. The CloudAnalyst simulator separates the user's global location into a few zones. Any region can be selected for the simulation, and the number of data centres can be adjusted correspondingly. The initial configuration setup will be the subject of the simulation.

### A. Modelling Technique

This algorithm works based on load sets assigned to virtual machine sets at a given time. The mathematical model is defined by the following equations to calculate the processing time for a task to be executed. Eq. (1) is given by  $n$  as the number of sets for the load ( $L$ ) or requests that need to be scheduled to servers.

$$L = \{L_1, L_2, \dots, L_n\} \quad (1)$$

Eq. (2) is given by  $k$  as the number of sets for virtual machines ( $V$ ) in a particular data centre ( $D$ ).

$$V = \{V_1, V_2, \dots, V_k\} \quad (2)$$

Eq. (3) is given by  $DL$  as the current data centre load.

$$DL = \{VL_1, VL_2, \dots, VL_k\} \quad (3)$$

Eq. (4) shows to find a function where the load set can be mapped (multiply) into virtual machine set ( $VL$ ) and  $f(L)$ , forming load  $VL_i$  of each virtual machine  $V_i$  to be essentially equal.

$$VL_1 \approx VL_2 \approx \dots \approx VL_k \quad (4)$$

Eq. (5) shows to calculate the time needed to allocate all tasks to virtual machine  $V_i$ , take  $\tau_0$  as the time to execute task  $L_0$ .

$$t_i = \sum_{0 \in f(L_i)(i=1, \dots, n)} \tau_0 \quad (5)$$

When there is only one virtual machine, which means  $k=1$  and all available tasks will be executed serially on that one virtual machine, Eq. (6) shows the execution time will be the summation of all time and can be calculated as  $T_1$ .

$$T_1 = \sum T_0 (0 = 1, \dots, n) \quad (6)$$

And when  $k>1$  which means there are more virtual machines available, all available tasks can be allocated (shared equally) to multiple servers. Making the serially executed task before becoming parallel since multiple virtual machines are working simultaneously. Thus, Eq. (7) shows the time needed to execute a task is calculated as  $T_k$ .

$$T_k = \max_{i=1, \dots, n} t_i \quad (7)$$

In conclusion, the goal is to solve the function to get the minimum of time needed to execute a task.

### III. METHODOLOGY

This section explains the details of the flow chart of the proposed study and the simulation for the analysis of Equally Spread Current Execution Load Modelling for Cloud Computing.

#### A. Flow Chart

Fig. 2 shows the ESCE algorithm flowchart which depicts how the algorithm handles user requests. To ensure equal loads for every virtual machine (VM) involved in the algorithm, ESCE uses the spread spectrum approach. Receiving a request, ESCE uses the spread spectrum approach. Receiving a request, the load balancer updates the index table with the VM status count and scans the request in the index table before allocating it to an available VM. The parallelogram represents the beginning of the user request, the square shape of the process carried out by the simulator to evaluate the request before sending it to the virtual machine, and the diamond is the choice made by the load balancer of which virtual machine to send the user request to be processed. The virtual machine count and availability are stored in data storage with a cylinder shape.

#### B. Simulation

Equally Spread Current Execution (ESCE) is a cloud analytic methodology that is used to optimize the performance of cloud services. It is a cost-effective method of achieving higher throughput in the cloud. ESCE works by dynamically spreading the current execution load across multiple servers in order to evenly distribute the load. This helps to reduce the load on any one server, thus increasing the overall performance of the cloud. Additionally, ESCE can also be used to improve the scalability of the cloud environment. By leveraging multiple servers, the cloud can handle larger workloads without experiencing any performance degradation. ESCE is an effective way to improve the performance and scalability of cloud services and can be implemented in a variety of cloud environments.

The simulation is started by identifying the desired system components such as virtual machines, storage, network components and the desired number of resources for each component. A simulation tool that will be used to measure the performance of the system was developed using CloudAnalyst. This tool should include the ability to measure system performance and resource utilization. The simulation began

with powering the software application. The user base (UB), data centre (DC), and load-balancing policy were identified in the simulation configuration, in this example, the Equally Spread Current Execution (ESCE). Execute the simulation tool and analyze the results. This includes measuring the average response time, resource utilization, throughput, and scalability of the system. The entire simulation lasted for 60 minutes. The number of data centres and virtual machines is a controlled variable in this simulation. The changes in outcome were determined by both variables. Based on a few initial setups, this software was used to simulate and study the cloud environment. The Equally Spread Current Execution (ESCE) load balancer paradigm was employed in this study to examine load-balancing performance. Response time, data centre processing time, and virtual machine (VM) cost are the outcomes tracked. The system configuration and workload have been optimized to achieve the desired performance and re-run the simulation tool and analyze the results. Fig. 3 illustrates how the CloudAnalyst simulator, an extension of the CloudSim toolkit, simulates scenarios based on internet behavior [40]. Build with a clear graphical user interface, the simulator's main objective is to separate experimentation activity from mapping exercises so the modeler may focus on the issues rather than the technicalities.

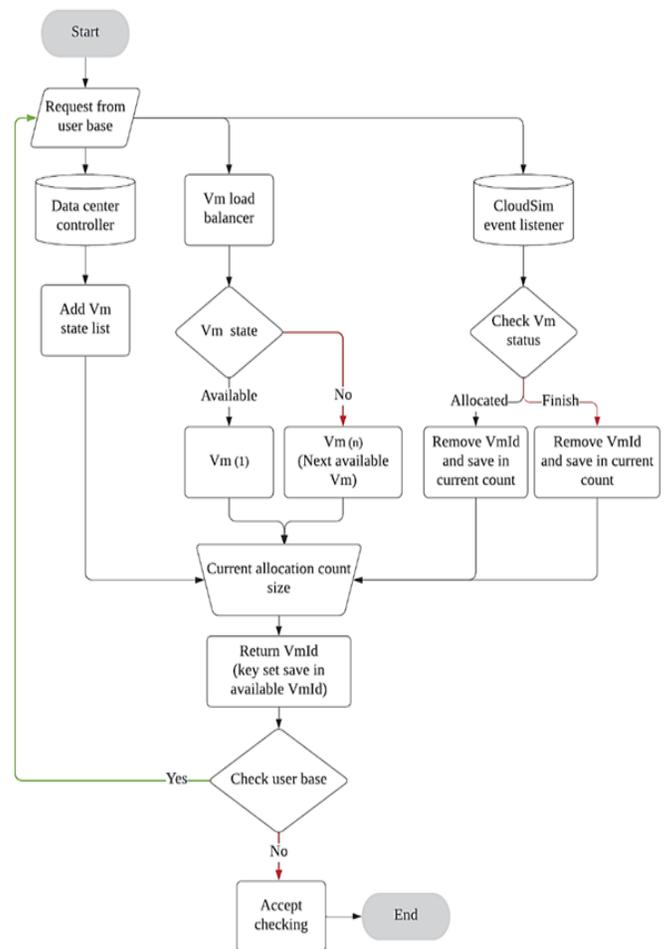


Fig. 2. ESCE execution process flowchart



Fig. 3. Graphical user interface (GUI) of CloudAnalyst

#### IV. RESULT AND DISCUSSIONS

This section describes the result of the simulation and the discussions for the simulation in the CloudAnalyst toolkit. This analysis was run in the CloudAnalyst simulator, the data is worked from an application “Facebook” known to be one of the examples of cloud computing applications. North America, Africa, and Oceania have been chosen to be the regions where user bases are assigned. The software identified these regions as R0, R4 and R5 respectively. There are more than these three regions in the software to be selected from, but they are chosen based on the approximate distance from each other to balance the outcome of simulation scenarios. The statistics of Facebook users in 2022 showed that R0 has 201.3 million users, R4 is the region with the highest number of users at 242.2 million and lastly, and R5 with 21.0 million.

##### A. Configurations

Table I presents the user number details for peak and off-peak hours. Table II presents the detailed settings for different simulation scenarios. There are a few different scenarios for simulation testing to get the desired results. Based on the statistical information, each region was set to only 10% of the total user number for peak-hour users and 5% for off-peak hours. The number of requests coming in is assumed to be once per five minutes for R0 and R5, while once per ten minutes for R4. The data size for a request is set to 100b. All simulation scenarios were run for 60 minutes with user grouping factor in user bases at 1000, request grouping factor in data centres at 100 and executable instructions length per request at 500 bytes. The load-balancing policy used is the Equally Spread Current Execution Load (ESCE) since it is the main objective of this analysis. The data centres are set to a memory of 1024 Mb. Other unmentioned configurations are unchanged in their default settings.

##### B. Response Time

Response time is the time taken to respond to requests coming in from clients. Lower response time means the load-balancing algorithm is doing a good job. Fig. 4 shows the analysis of the DC location as the variable. Three different simulation scenarios run with one DC in each simulation with the DC being located at R0, R4 and R5 alternately. Average overall response time showed DC located at R0 having the shortest response time compared to DC located at R4 and R5.

TABLE I. User NUMBER FOR PEAK AND OFF-PEAK HOURS

Region	User number	
	Peak Hour(million)	Off-Peak Hour (million)
R0	4.026	2.013
R4	4.844	2.422
R5	0.420	0.210

TABLE II. DETAILED SETTINGS FOR DIFFERENT SCENARIOS

Simulation	DC Number	Different scenarios		
		VM Number	DC Location (Region)	Service Broker Policy
1	1	100	R0	Closest Data Centre
2	1	100	R4	Closest Data Centre
3	1	100	R5	Closest Data Centre
4	1	1000	R0	Closest Data Centre
5	1	10	R0	Closest Data Centre
6	3	100	R0, R4, R5	Closest Data Centre
7	3	100	R0, R4, R5	Optimise Response Time
8	3	100	R0, R4, R5	Reconfigure Dynamically with Load



Fig. 4. Average overall response time analysis based on one data centre (a) at a different location

Fig. 5 shows the result of the simulation run with three different numbers of VMs. 10 VMs gives the shortest response time analysis at 13.05 seconds among all three followed by 100 VMs at 15.05 seconds and 1000 VMs at 15.25 seconds. 100 and 1000 VMs gave only slightly different result from each other but is still significant with the difference being approximately 0.2 seconds which is quite big since original data are analyzed in milliseconds.

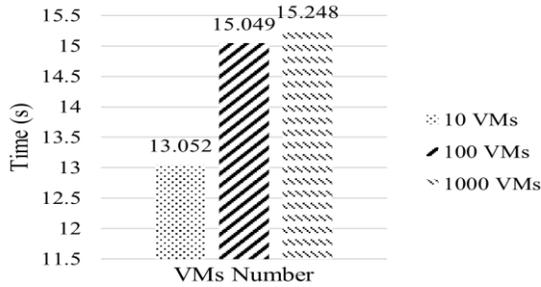


Fig. 5. Average overall response time analysis based on one data centre at a different number of VM

Fig. 6 presents the result given by applying different brokerage policies in three different simulations to analyze the response time given by each brokerage policy. The Optimize Response Time Brokerage policy has the shortest response time at 8.63 seconds, followed by reconfigure dynamically with load at 13.31 seconds response time and the closest data centre at 13.40 seconds. Response time is also observed by region and all three regions gave different results depending on the variables carried out for each simulation. Fig. 7 illustrates the result of response time with different DC locations. This result showed that all three locations gave R5 the shortest response time.

Fig. 8 presents all VMs number categories that had the same result with R5 being the region with the shortest average response time. Fig. 9 shows that given three DC in total with one at each region respectively, R5 also have the shortest response time among the three regions which is small compared in number.

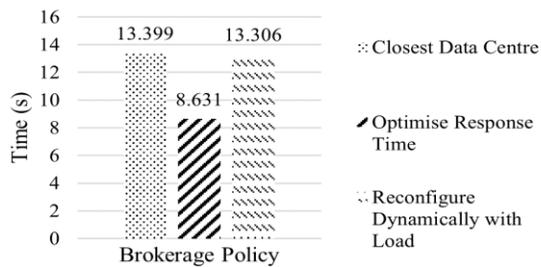


Fig. 6. Average overall response time analysis based on three data centres for different brokerage policy

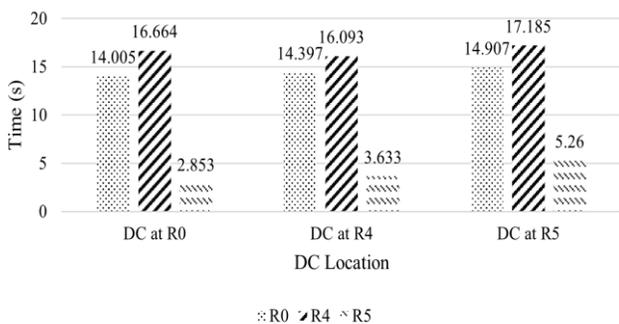


Fig. 7. Average response time by region based on one data centre at a different location

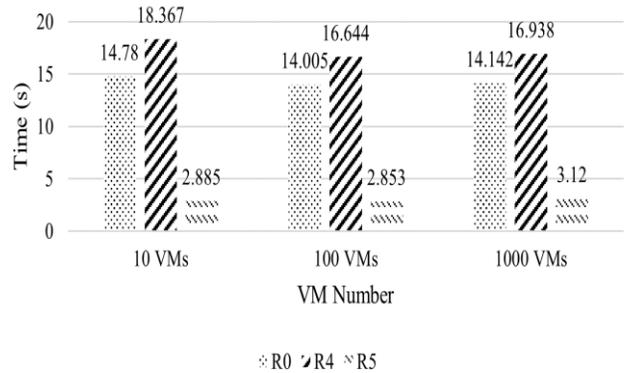


Fig. 8. Average response time analysis by region based on one data centre with a different number of VM

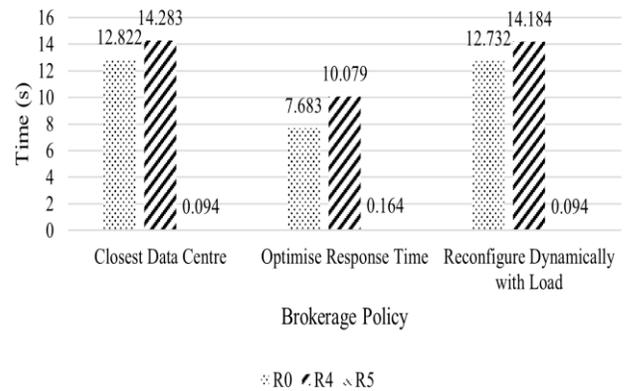


Fig. 9. Average response time analysis by region based on three data centres for different brokerage policies

### C. Average Data Centre Processing Time

The data centre processing time analyzed the DC capability to work under different conditions. Fig. 10 shows one DC placed at different locations giving the shortest average processing time of 14.55 seconds when located at R4. Fig. 11 shows processing time took only 12.38 seconds for 10 VMs in one DC compared to 14.71 seconds for 100 VMs and 14.90 seconds for 1000 VMs.



Fig. 10. Average DC processing time analysis based on one data centre at a different DC location

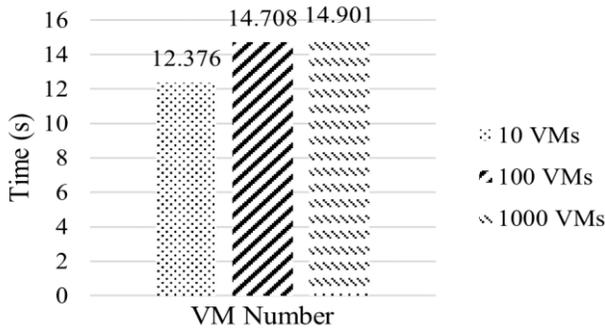


Fig. 11. Average DC processing time analysis based on one data centre with a different number of VM

Fig. 12 describes the result of DC processing time for the simulation with different brokerage policies. Results showed that Optimize Response Time Brokerage policy gave the shortest processing time for the data centre at 8.36 seconds compared to the other two brokerage policies.

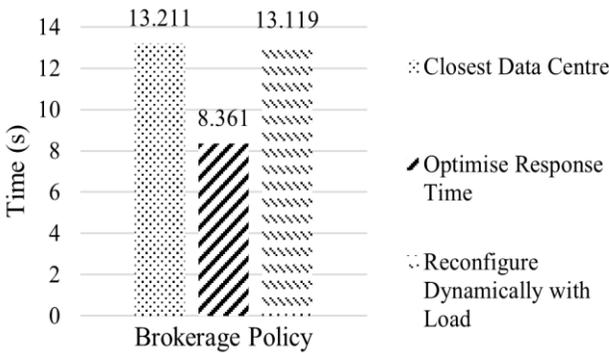


Fig. 12. Average DC processing time analysis based on three data centres for different brokerage policy

**D. Resource Utilization (Cost)**

Defined to check the utilization of resources. It is related to cost. Resource utilization in a system should be maximized to avoid clients paying for any unused resources. All cost used in the simulation is as shown in Table III.

One data centre with 100 VMs would cost approximately 10\$ each. Brokerage policy simulation only used one data centre for each different brokerage policy. Fig. 13 shows that all cost does not differ much from each other.

TABLE III. COST-DETAILED SETTING FOR EACH DC

Category	Price (\$)
Cost per VM (\$/hr)	0.1
Memory Cost (\$/s)	0.05
Storage cost (\$/s)	0.1

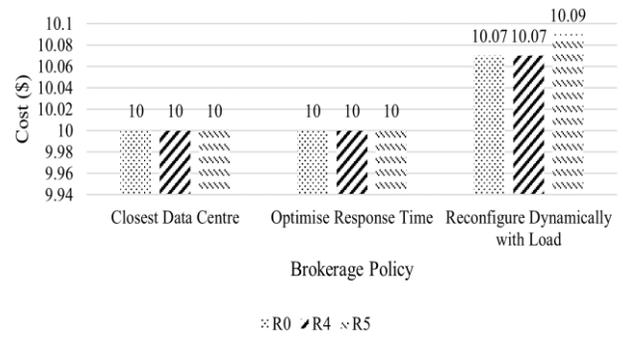


Fig. 13. VM cost analysis by region based on three data centres for different brokerage policy

VM cost would differ depending on how much VM is assigned in a data centre. Fig. 14 shows the result for the simulation of different DC locations but all having one data centre at a time and 100 VMs. Each DC cost the same since all have the same VM number. Fig. 15 shows a clear difference since the three simulation uses a different amount of VM in one data centre. 10 VMs is the cheapest since 1 VM used per hour only cost 0.1\$. 1000 VMs cost only 80\$ and not 100\$ since normally cloud plan offers some discount for certain packages subscribed.

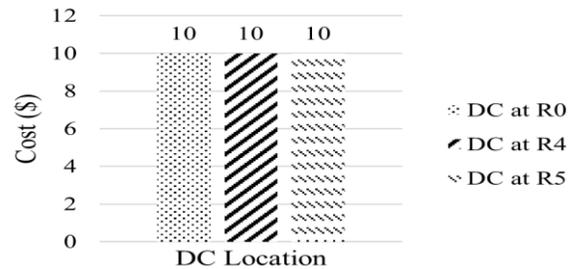


Fig. 14. VM cost analysis based on one data centre at a different DC location

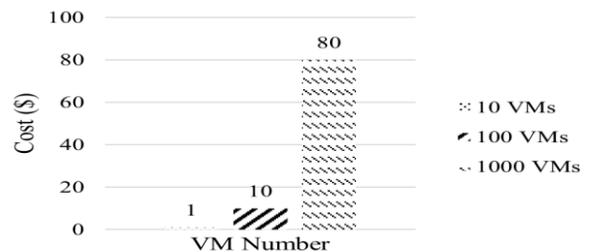


Fig. 15. VM cost analysis based on one data centre with a different number of VM

Fig. 16 shows three data centres would cost approximately 30\$ since one data centre costs 10\$. The reconfiguring dynamically with load brokerage policy gave off 30.22\$ since the machine worked according to load changes. The 0.22\$ difference is the result of algorithm work to be reconfigured according to load.

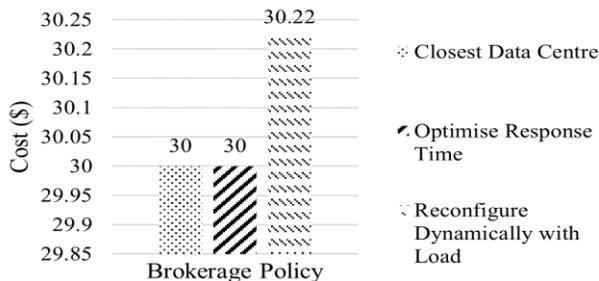


Fig. 16. VM cost analysis based on three data centres for different brokerage policy

Data transfer costs depended on the DC location. Data transfer costs would be higher if the distance between the DC location and the region where requests come from is quite far. Fig. 17 shows that in the closest data centre and reconfigure dynamically with load brokerage policy the region with the cheapest data transfer cost is R5 since the DC location is closer to any of the other two regions but, in optimizing Response Time Brokerage policy, the algorithm had to work based on the policy despite the location of DC. Hence, the cheapest cost for this brokerage policy only went as low as 106.91\$ at R0. The total data transfer cost for one data centre at a different location is lowest when DC is located at R0 and R5 with both being 385.11\$. Data transfer cost is calculated at a total of DC transferring data to all assigned regions. Fig. 18 illustrates there is only one data centre at a time which would make all three regions assigned to the one DC to carry out task execution work.

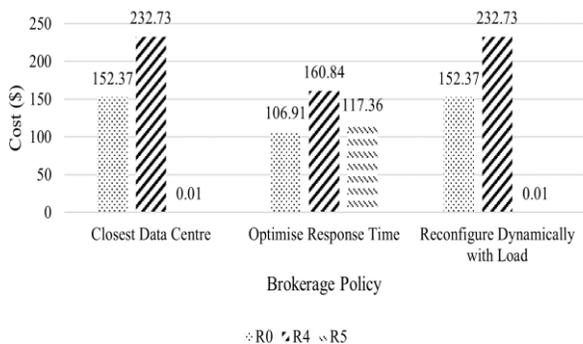


Fig. 17. DT cost analysis by region based on (a) three data centres for different brokerage policy

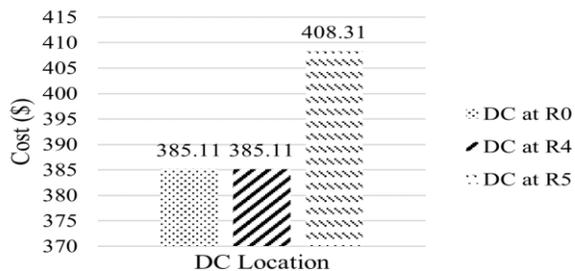


Fig. 18. DT cost analysis by region based on one data centre at a different DC location

Fig. 19 shows 100 VMs and 1000 VMs having the same results of 385.11\$ while 10 VMs costs 504.3\$. This is because the lower number of VMs had to do more work to execute data before transferring it to the user in their respective regions.

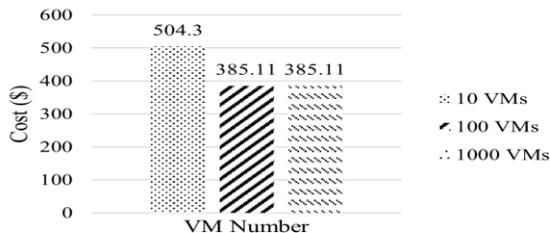


Fig. 19. DT cost analysis based on one data centre with a different number of VM

Fig. 20 illustrates the brokerage policies have equal same results of 385.11\$ in data transfer cost since simulations were carried out with one data centre located respectively at each region. All regions have their own data centre to process tasks.

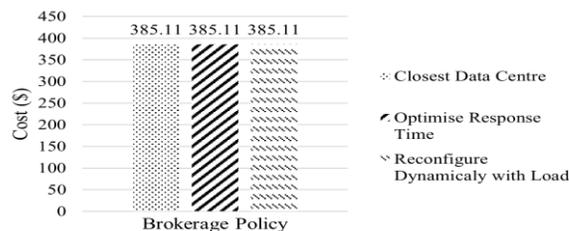


Fig. 20. DT cost analysis by region based on three data centres for a different brokerage policy

## V. CONCLUSIONS

This study has successfully analysed the three main factors of cloud computing which are off-peak resource utilisation, minimum data processing time, and minimal average response time. To optimize the response time brokerage policy for cloud computing, a simulation for the equally spread current execution load has been developed. This simulation has considered the current workload of each cloud system, the expected workload in the near future, and the existing resources of each system. The goal of the model is to ensure that the workload is evenly balanced across the cloud systems, to ensure that no one system is overloaded while another is underutilized. The result of this simulation was able to allocate the workload among the cloud systems in a way that optimizes response time and meets the goals of the system. This can be done by assigning the workload to the system that can best handle it, or by using techniques such as load-balancing or resource scheduling. It should also consider the future growth of the cloud system. This can be done by allowing for future increases in workloads.

## ACKNOWLEDGMENT

Authors would like to thank Research Management Centre (RMC), Universiti Teknologi MARA, Shah Alam, Selangor for the support funding in this research from grant no. 600-RMC/GPMST5/3(038/2021) and Institute for Big Data

Analytics and Artificial Intelligence (IBDAAI) in supporting this research.

## REFERENCES

- [1] T. Azmat, D. Kumar, and V. K. Dwivedi, "A Novel Approach towards an Efficient Load-balancing Algorithm in Cloud Computing," in 2019 4th International Conference on Information Systems and Computer Networks, ISCON 2019, 2019, pp. 572-577, doi: 10.1109/ISCON47742.2019.9036309.
- [2] A. S. M. Al-Obaidi, A. Al-Qassar, A. R. Nasser, A. Alkhayat, A. J. Humaidi, and I. K. Ibraheem, "Embedded design and implementation of mobile robot for surveillance applications," Indonesian Journal of Science and Technology, Article vol. 6, no. 2, pp. 427-440, 2021, doi: 10.17509/ijost.v2i2.
- [3] L. Hasanah, W. L. Hakim, A. Aminudin, S. K. Sahari, and B. Mulyanti, "A design and performance analysis of a telemetry system for remote monitoring of turbidity of water during the covid-19 pandemic," Indonesian Journal of Science and Technology, Article vol. 5, no. 2, pp. 299-307, 2020, doi: 10.17509/ijost.v5i2.24705.
- [4] Saripudin, A. Djohar, D. Rohendi, and A. G. Abdullah, "Developing information technology in opencourseware: From movements to opportunities in Asia," Indonesian Journal of Science and Technology, Article vol. 5, no. 3, pp. 308-320, 2020, doi: 10.17509/ijost.v5i3.24886.
- [5] K. Stanoevska-Slabeva and T. Wozniak, "Cloud basics-An introduction to cloud computing," in Grid and Cloud Computing: A Business Perspective on Technology and Applications, 2010, pp. 47-61.
- [6] S. M. Shetty and S. Shetty, "Analysis of load-balancing in cloud data centers," Journal of Ambient Intelligence and Humanized Computing, pp. 1-9, 2019, doi: 10.1007/s12652-018-1106-7.
- [7] N. S. M. Pakhrudin, M. Kassim, and A. Idris, "A review on orchestration distributed systems for IoT smart services in fog computing," International Journal of Electrical and Computer Engineering, vol. 11, no. 2, p. 1812, 2021. [Online]. Available: 10.11591/ijece.v11i2.pp1812-1822.
- [8] V. Andrikopoulos, T. Binz, F. Leymann, and S. Strauch, "How to adapt applications for the Cloud environment: Challenges and solutions in migrating applications to the Cloud," Computing, Article vol. 95, no. 6, pp. 493-535, 2013, doi: 10.1007/s00607-012-0248-2.
- [9] T. Diaby and B. B. Rad, "Cloud Computing: A review of the Concepts and Deployment Models," International Journal of Information Technology and Computer Science, vol. 9, pp. 50-58, 2017, doi: 10.5815/IJITCS.2017.06.07.
- [10] A. Mishra and D. Tiwari, "A Proficient Load-balancing Using Priority Algorithm in Cloud Computing," in Proceedings of the 2020 IEEE International Conference on Machine Learning and Applied Network Technologies, ICMLANT 2020, 2020, doi: 10.1109/ICMLANT50963.2020.9355972.
- [11] M. Hamdani, Y. Aklouf, and H. Chaalal, "A Comparative Study on Load-balancing Algorithms in Cloud Environment," in ACM International Conference Proceeding Series, 2020, doi: 10.1145/3447568.3448466.
- [12] V. S. Handur, S. Belkar, S. Deshpande, and P. R. Marakumbi, "Study of load-balancing algorithms for Cloud Computing," in Proceedings of the 2nd International Conference on Green Computing and Internet of Things, ICGCIoT 2018, 2018, pp. 173-176, doi: 10.1109/ICGCIoT.2018.8753091.
- [13] A. Singh and R. Kumar, "Performance evaluation of load-balancing algorithms using cloud analyst," in Proceedings of the Confluence 2020 - 10th International Conference on Cloud Computing, Data Science and Engineering, 2020, pp. 156-162, doi: 10.1109/Confluence47617.2020.9058017.
- [14] H. Rai, S. K. Ojha, and A. Nazarov, "Cloud Load-balancing Algorithm," in Proceedings - IEEE 2020 2nd International Conference on Advances in Computing, Communication Control and Networking, ICACCCN 2020, 2020, pp. 861-865, doi: 10.1109/ICACCCN51052.2020.9362810.
- [15] A. Markandey, P. Dhamdhere, and Y. Gajmal, "Data access security in cloud computing: A review," in 2018 International Conference on Computing, Power and Communication Technologies, GUCON 2018, 2019, pp. 633-636, doi: 10.1109/GUCON.2018.8675033.
- [16] I. N. Falisha, T. W. Purboyo, and R. Latuconsina, "Experimental Model for Load-balancing in Cloud Computing Using Equally Spread Current Execution Load Algorithm," International Journal of Applied Engineering Research, vol. 13, no. 2, pp. 1134-1138, 2018, doi: 10.37622/000000.
- [17] K. Kaur and R. Mahajan, "Equally spread current execution load algorithm-a novel approach for improving data centre's performance in cloud computing," International Journal on Future Revolution in Computer Science & Communication Engineering, vol. 4, no. 8, pp. 08-10-08-10, 2018.
- [18] C. Xue, "Method and implementation of server load-balancing in cloud computing," in Proceedings - 2018 3rd International Conference on Mechanical, Control and Computer Engineering, ICMCCE 2018, 2018, pp. 511-513, doi: 10.1109/ICMCCE.2018.00113.
- [19] F. Alhaidari and T. Z. Balharith, "Enhanced Round-Robin Algorithm in the Cloud Computing Environment for Optimal Task Scheduling," Computers, vol. 10, no. 5, p. 63, 2021. doi: https://doi.org/10.3390/computers10050063.
- [20] S. Kumar and A. Dumka, "Load-balancing with the Help of Round Robin and Shortest Job First Scheduling Algorithm in Cloud Computing," in Proceedings of International Conference on Machine Intelligence and Data Science Applications, Singapore, M. Prateek, T. P. Singh, T. Choudhury, H. M. Pandey, and N. Gia Nhu, Eds., 2021// 2021: Springer Singapore, pp. 213-223. [Online]. Available: https://doi.org/10.1007/978-981-33-4087-9\_19. [Online]. Available: https://doi.org/10.1007/978-981-33-4087-9\_19.
- [21] S. I. Suliman et al., "An effective energy-efficient virtual machine placement using clonal selection algorithm," International Journal of Advanced Technology and Engineering Exploration, vol. 8, no. 75, p. 412, 2021. doi: 10.19101/IJATEE.2020.762129.
- [22] S. Kaur and T. Sharma, "Efficient load-balancing using improved central load-balancing technique," in 2018 2nd International Conference on Inventive Systems and Control (ICISC), 19-20 Jan. 2018 2018, pp. 1-5, doi: 10.1109/icisc.2018.8398857.
- [23] V. Joshi, "Load-balancing Algorithms in Cloud Computing," (in English), International Journal of Research in Engineering and Innovation, vol. 3, pp. 530 - 532, 2019-12-14 2019.
- [24] T. Deepa and D. Cheelu, "A comparative study of static and dynamic load-balancing algorithms in cloud computing," in 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS), 1-2 Aug. 2017 2017, pp. 3375-3378, doi: 10.1109/icecds.2017.8390086. [Online]. Available: 10.1109/icecds.2017.8390086.
- [25] H. Ye, J. Guo, and X. Li, "Delay-Aware and Profit-Maximizing Task Migration for the Cloudlet Federation," International Journal of Advanced Computer Science and Applications, Article vol. 13, no. 10, pp. 420-428, 2022, doi: 10.14569/ijacsa.2022.0131050.
- [26] R. Kamran, A. A. El-Moursy, and A. Abdelsamea, "Efficient HPC and Energy-Aware Proactive Dynamic VM Consolidation in Cloud Computing," International Journal of Advanced Computer Science and Applications, Article vol. 13, no. 10, pp. 858-869, 2022, doi: 10.14569/ijacsa.2022.01310102.
- [27] R. Alanazi, "Analysis of Privacy and Security Challenges in e-Health Clouds," International Journal of Advanced Computer Science and Applications, Article vol. 13, no. 9, pp. 484-489, 2022, doi: 10.14569/ijacsa.2022.0130955.
- [28] Y. Qiu, B. Sun, Q. Dang, C. Du, and N. Li, "Fine-grained Access Control Method for Blockchain Data Sharing based on Cloud Platform Big Data," International Journal of Advanced Computer Science and Applications, Article vol. 13, no. 10, pp. 24-31, 2022, doi: 10.14569/ijacsa.2022.0131004.
- [29] T. Saxena and V. Chourey, "A survey paper on cloud security issues and challenges," in Proceedings of the 2014 Conference on IT in Business, Industry and Government: An International Conference by CSI on Big Data, CSIBIG 2014, 2014, doi: 10.1109/CSIBIG.2014.7056957.
- [30] M. Aziz, "Advanced green technologies toward future sustainable energy systems," Indonesian Journal of Science and Technology, Article vol. 4, no. 1, pp. 89-96, 2019, doi: 10.17509/ijost.v4i1.15805.

- [31] M. Mesbahi and A. M. Rahmani, "Load-balancing in cloud computing: a state of the art survey," *Int. J. Mod. Educ. Comput. Sci.*, vol. 8, no. 3, p. 64, 2016, doi: 10.5815/ijmecs.2016.03.08.
- [32] T. Nakrani, D. Hiran, C. Sindhi, and M. I. Sandhi. Genetic Algorithm Based Task Scheduling for Load-balancing in Cloud, *Lecture Notes on Data Engineering and Communications Technologies*, vol. 52, pp. 283-293, 2021.
- [33] S. Handur Vidya and R. Marakumbi Prakash, "Response time analysis of dynamic load-balancing algorithms in Cloud Computing," in *Proceedings of the World Conference on Smart Trends in Systems, Security and Sustainability, WS4 2020*, 2020, pp. 371-375, doi: 10.1109/WorldS450073.2020.9210305.
- [34] A. Kaur, B. Kaur, and D. Singh, "Optimization techniques for resource provisioning and load-balancing in cloud environment: a review," *International Journal of Information Engineering and Electronic Business*, vol. 9, no. 1, p. 28, 2017, doi: 10.5815/ijieeb.2017.01.04.
- [35] M. Hashemi and A. Masoud, "Load-balancing Algorithms in Cloud Computing Analysis and Performance Evaluation," *IEEE*, vol. 3, no. 4, 2020.
- [36] A. Y. Ahmad and A. Y. Hammo, "A Comparative Study of the Performance of Load-balancing Algorithms Using Cloud Analyst," *Webology*, vol. 19, no. 1, 2022, doi: 10.14704/WEB/V19I1/WEB19328.
- [37] S. Lamba and D. Kumar, "A comparative study on load-balancing algorithms with different service broker policies in cloud computing," *International Journal of Computer Science and Information Technologies*, vol. 5, no. 4, pp. 5671-5677, 2014.
- [38] S. Mohapatra, K. S. Rekha, and S. Mohanty, "A comparison of four popular heuristics for load-balancing of virtual machines in cloud computing," *International Journal of Computer Applications*, vol. 68, no. 6, 2013, doi: 10.5120/11586-6922.
- [39] J. Rathore, B. Keswani, and V. S. Rathore. Analysis of Load-balancing Algorithms Using Cloud Analyst, *Advances in Intelligent Systems and Computing*, vol. 841, pp. 291-298, 2019.
- [40] B. Wickremasinghe, R. N. Calheiros, and R. Buyya, "CloudAnalyst: A cloudsim-based visual modeller for analysing cloud computing environments and applications," in *Proceedings - International Conference on Advanced Information Networking and Applications, AINA*, 2010, pp. 446-452, doi: 10.1109/AINA.2010.32.

# Research on the Derivative Rule and Estimation Methods of Intelligent High-Speed Railway Investment Estimation

Yang Meng, Chuncheng Meng, Xiaochen Duan\*

College of Management, Shijiazhuang Tiedao University, Shijiazhuang, Hebei 050043, China

**Abstract**—Taking the investment estimation of intelligent construction of high-speed railway as the research object, based on the historical data of investment of similar high-speed railway projects, this paper builds an explanatory structure model, establishes a system dynamic (SD) model of investment estimation of intelligent construction of high-speed railway, and puts forward suggestions for supplementing the labor value theory and improving the value-added tax. The paper carries out in-depth research and analysis in the following aspects: 1) the list of influencing factors for investment estimation of intelligent construction of high-speed railway in the feasibility study stage is constructed, and the interpretative structural model (ISM) is constructed to sort out the relationship between the influencing factors; 2) the SD model of intelligent construction cost estimation of high-speed railway is established to improve the accuracy of investment estimation of intelligent high-speed railway construction; 3) put forward suggestions and schemes for improving investment estimation content of intelligent construction of high-speed railway under high intelligence; 4) improve and supplement the labor value theory and the value-added tax base.

**Keywords**—High-speed railway; intelligent construction; investment estimation; interpretative structural model; system dynamics

## I. INTRODUCTION

The level of intelligence in the future will largely affect the competitiveness of countries in the field of high-speed railroads, which brings pressure on the intelligent transformation of China's high-speed railroads while providing a huge opportunity for development. However, with the rapid development of high-speed railway construction under high intelligence, the human labor required in the construction process of high-speed railway is becoming less and less, and "machine instead of human" is becoming more and more common at the construction site; in the near future, less humanized and unmanned will become the mainstream of high-speed railway engineering construction [1]. The corresponding content and method of investment estimation of high-speed railway construction under high intelligence are only suitable for the low and medium intelligence level. With the enhancement of intelligence, the labor cost and management cost are greatly reduced and the robot cost increases, i.e., the content and structure of investment estimation is undergoing a fundamental change. The value-added tax (VAT) calculation theory based on the labor theory of value, where human labor creates new value, is under attack.

Therefore, it is important to analyze and study the characteristics of the factors influencing the investment estimation of intelligent construction of high-speed railroads and their mechanism of action, construct an investment estimation system for intelligent construction of high-speed railroads based on system dynamics, establish a case empirical system to accompany the system for empirical analysis, and propose improvements to the labor value theory and VAT billing base.

With France, Germany and Japan as the prominent examples, all countries in the world have launched research on intelligent railroad construction [2]. German railroads aim to achieve information management and widely promote building information modeling (BIM) technology. In 2014, it proposed to apply 5D BIM technology to the whole process management of German railroad construction and started the intelligent change with the iTWO 5D BIM platforms, such as Bahnhof Horrem station, Rastatter tunnel, Filstal bridge, etc [3, 4]. France, one of the traditional railroad powers, now also places the development of railroad information technology in its development strategy, for example, the digital French railway strategy officially launched in 2015, which strives to focus on the construction of Internet for railroad stations, trains, and road networks and plans to build a railroad system with convenient transportation, strong competitiveness, and close relevance to future transportation from 2031 to 2040 [5, 6]. The United Kingdom expects significant cost reductions in digital signaling, intelligent infrastructure and train control after 2027 and takes applications of intelligent technologies as a daily means of operating the railway network [7]. Japan has taken the lead in the development of intelligence in the field of construction and has started the construction of intelligent coverage of public works management ten years ago to successfully realize the paperlessness of all information in the whole life cycle of construction, which has greatly improved the efficiency of construction and also reduced construction costs [8].

In recent years, Chinese scholars have carried out many research works on the application of system dynamics in the field of engineering construction. Zhou constructed a model of project cost management system based on system dynamics to realize the dynamic evaluation of project costs [9]. Yue uses the system dynamics method to build the project quality system model to realize the quality management of construction projects [10]. Mao constructed a system dynamics model based

on the framework of “BIM-cooperative subject”, pointed out the importance and relationship of relevant influencing factors, and guided the development direction for proposing the strategy of synergy effect of cooperative subject [11]. Gao constructed a risk model for engineering projects based on system dynamics theory and verified it with petroleum projects, which provides a development direction for engineering project risk management [12]. By establishing a system dynamics model for resource allocation of engineering projects, Zhong et al. proposed that the resource allocation strategy should be formulated based on the matching between the characteristics of project operation mode and multiple resources [13]. Based on the system dynamics model, Liu et al. studied the relationship between engineering schedule and cost management [14]. Chen et al. constructed a system dynamics model to study the logical relationship between the factors influencing the cost of prefabricated building and selected rational measures to optimize the cost, improving the current situation of cost management of prefabricated construction projects [15].

In summary, many scholars have used system dynamics to study engineering investment estimation, but most of the studies focus on tunnel investment, component cost, or estimation of part of the cost of the overall project such as material cost, and the system dynamics-based intelligent construction investment estimation for high-speed railway is currently not reported in the literature. This paper explored the method of intelligent construction investment estimation for high-speed railways, identified the factors influencing intelligent construction investment estimation for high-speed railways, constructed an interpretative structural model (ISM) model to explore the interactions between the factors, and built a system dynamic (SD) model. Moreover, SD, scenario simulation, and case inference were used to study the trend of various cost changes of intelligent construction of high-speed railway under high intelligence (70-100%). According to the evolution trend, the corresponding investment estimation cost content and rate improvement methods were proposed to provide academic theoretical basis and support for the investment estimation decision of high-speed railway under high intelligence.

This paper took the intelligent construction investment estimation of high-speed railway as the research subject, constructed the ISM to sort out the relationship of influencing factors of intelligent construction investment estimation of high-speed railway; on this basis, the SD model of intelligent construction investment estimation of high-speed railway was constructed to achieve high-precision estimation; finally, intelligent construction cost correction scheme and tax rate increase correction scheme were proposed under high intelligence (70%-100%).

## II. RELATED BASIC THEORIES

### A. Interpretative Structural Model Theory

The interpretative structural model (ISM) often presents the results of analytic hierarchy process in a skeleton diagram, which clearly and intuitively illustrates the role of every element in the interaction relationship, and it shows obvious advantages in sorting out the complex, multi-layered factor

relationship [16]. Considering the intelligent construction investment estimation of high-speed railway as a whole, the ISM can quickly sort out the relationship between its influencing factors and thus find the key influencing factors. The specific steps are shown below.

Firstly, the correlation relationship between factors is determined, and an adjacency matrix is constructed. The interaction relationship between influencing factors is analyzed. If factor  $S_i$  has influence on factor  $S_j$ , then it is stipulated that there is a direct influence relationship between them; otherwise, it is considered that there is no direct influence relationship between them.

The next step is to determine the recursive relationship between factors in two steps. The first step is to calculate the reachable matrix, and the other step is to determine the level division of different factors.

1) *Find the reachable matrix.* A reachable matrix describes the extent to which the factors can be reached after a certain length of pathway between them through the matrix form, which often requires the indirect role of intermediate factors. Boolean matrix operation rules make the following provision:

when  $(A + I)^{K-1} \neq (A + I)^K = (A + I)^{K+1}$ ,  $(A + I)^{K-1}$  is the desired reachable matrix M.

2) *Divide the hierarchical structure.* Based on the obtained reachable matrix M, the reachable set and the prior set of the factors are obtained, and the influencing factors are divided into different hierarchies.  $R(S_i)$  represents the reachable set, which is the set of all factors that can be reached from  $S_i$ , i.e., the set of factors that can be directly influenced by  $S_i$ ;  $A(S_i)$  represents the prior set, which is the set of all factors that can reach  $S_i$ , i.e., the set of factors that have direct influence on it. Based on this, the intersection set  $C(S_i)$  of the them is obtained.

When the reachable set and the intersection set contain exactly the same influencing factors, they are classified as influencing factors at the same layer. Based on this, the rows and columns where such factors are located are removed, and the remaining influencing factors are classified again by the above method, and so on until all factors are stratified.

### B. System Dynamics Model

SD is a highly dynamic scientific method of analysis based on the whole process of a system [17], which can solve complex and nonlinear systemic problems. The main parameters of the model are divided into four categories: constants, initial values, linear functions, and table functions.

1) Estimation of the constants and initial values with the GM(1,1) model.

Step 1: calculate the cumulative generation sequence.

$$x^{(1)}(k) = \sum_{i=1}^k x^{(0)}(i), k=1, 2, \dots, n \quad (1)$$

Step 2: find the mean series.

$$z^{(1)}(k) = \frac{x^{(1)}(k) + x^{(1)}(k+1)}{2}, k = 1, 2, \dots, n-1 \quad (2)$$

Step 3: calculate intermediate parameters  $C, D, E, F$ .

$$C = \sum_{k=2}^n z^{(1)}(k), D = \sum_{k=2}^n x^{(0)}(k), E = \sum_{k=2}^n z^{(1)}(k) * x^{(0)}(k), F = \sum_{k=2}^n z^{(1)}(k) \frac{\sum_{i=1}^n \sum_{j=1}^m q_j x_{ij}}{\sqrt{(\sum_{i=1}^n \sum_{j=1}^m p_j x_{ij})^2 + (\sum_{i=1}^n \sum_{j=1}^m q_j x_{ij})^2}} \quad (12)$$

Step 4: development coefficient  $a$  and grey action quantity coefficient  $b$ .

$$a = \frac{CD - (n-1)E}{(n-1)F - C^2} \quad (4)$$

$$b = \frac{DF - CE}{(n-1)F - C^2} \quad (5)$$

2) Estimation of the values of the factors of the linear functional relationship.

Step 1: determine the subjective weights using the G1 method.

It is assumed that the impact evaluation index system for investment estimation of high-speed railway construction projects has  $n$  evaluation indexes,  $\{x_1, x_2, x_3, \dots, x_n\}$  and  $w_i$  is the subjective weight of the  $i$ -th indicator. The expert survey method determines that these  $n$  elements have the following relationships:

$$x_1 \geq x_2 \geq x_3 \geq \dots \geq x_n \quad (6)$$

The ratio of the importance degree of evaluation indicator  $x_{i-1}$  to  $x_i$  is  $w_{i-1} / w_i$ , denoted as:

$$r_i = \frac{w_{i-1}}{w_i} (k = n, n-1, \dots, 3, 2). \quad (7)$$

Then, the subjective weights are calculated:

$$w_n = (1 + \sum_{i=2}^n (\prod_{k=i}^n r_k))^{-1} \quad (8)$$

$$w_{i-1} = r_i w_i (k = n, n-1, \dots, 3, 2) \quad (9)$$

Step 2: the entropy method is used to determine the objective weight.

The characteristic weight of the  $i$ -th expert under the  $j$ -th indicator is calculated:

$$p_{ij} = \frac{x_j}{\sum_{i=1}^m x_{ij}} \quad (10)$$

Step 3: the comprehensive weighting method determines the final weight.

The subjective weight is  $p_j$ , and the objective weight is  $q_j$ .

$$k_1 = \frac{\sum_{i=1}^n \sum_{j=1}^m p_j x_{ij}}{\sqrt{(\sum_{i=1}^n \sum_{j=1}^m p_j x_{ij})^2 + (\sum_{i=1}^n \sum_{j=1}^m q_j x_{ij})^2}} \quad (11)$$

The weight coefficient is substituted into the formula:

$$w_j = k_1 p_j + k_2 q_j \quad (13)$$

The final weight is obtained.

3) Estimation of the values of the factors of the table function relationship.

In the whole system flow chart, there is not only a linear function relationship between the factors, but also a table function relationship, for example, the relationship between material dynamic management capability and management intelligence level and the relationship between material dynamic management capability and material unit price influence factor.

### III. MODEL CONSTRUCTION

#### A. Construction Ideas

The intelligent construction system of high-speed railway is a dynamic and complex large system, involving many influencing factors, and the influencing factors are complex, stochastic, dynamic, and coupled, making the intelligent construction investment of high-speed railway show time-series non-linear dependent variable characteristics, such as non-linear, stochastic, dynamic, changeable, and prominent. Moreover, the intelligent construction system of high-speed railway in China is at the world leading level, so there are less historical data to draw on. For these reasons, this thesis used structural explanatory equations, historical data, literature questionnaires, and expert interviews to mine the factors influencing the intelligent construction investment of high-speed railways, used case inference and SD methods to analyze the evolutionary trends and derived mechanisms of the intelligent construction investment estimation, and constructed the corresponding investment estimation prediction and improvement model to effectively improve the accuracy and reliability of the intelligent construction investment estimation.

#### B. Model Composition

1) Influencing factor structure explanatory equation identification module.

a) Preliminary identification based on literature analysis method: We obtained more than 200 papers by searching the keywords of "intelligent high-speed railway", "intelligent construction", "investment estimation factors", and "investment estimation index". Twenty-three papers that cover a wide range of factors and were included in recent time were screened according to the criteria of citation frequency and the grade of journals and selected as the basis for identification to summarize the influencing factors of investment estimation of intelligent construction of high-speed railways.

b) *Additional identification based on the case study method:* To supplement the impact of intelligence on investment estimation, this paper took Beijing-Zhangjiakou and Beijing-Xiongan high-speed railways and three typical intelligent construction projects with large difficulties in the Zhengzhou-Wanzhou high-speed railway as examples to sort out the influencing factors of high-speed railway investment estimation.

c) *List of factors influencing investment estimation for intelligent construction of high-speed railways:* According to the feedback results of the expert pilot survey, the initial set of influencing factors was determined. Seventeen influencing factors were finally determined after questionnaire survey and analyzing and processing the data with SPSS software, including labor, material and machine costs, management costs, design depth, management efficiency, degree of economic development in the region, traditional construction machine configuration, intelligent robot configuration, construction site management level, dynamic material management capability, professional quality of management personnel, project complexity, resource allocation efficiency, intelligent degree of geological exploration, intelligent degree of construction, degree of equipment mechanization, information construction level, and intelligent degree of management.

2) Analysis of the evolutionary trend and derivative mechanism of the impact of influencing factors on investment Intelligent construction investment estimation for high-speed railways was viewed as a holistic system in which every influencing factor interacts with each other. A matrix was constructed for the influencing factors of the intelligent high-speed highway construction investment estimation using an ISM algorithm to quickly sorted out the direct or indirect relationships between the influencing factors [18].

a) *Determine the correlation between factors and construct the adjacency matrix:* The influencing factors were considered as a system consisting of 17 elements. An adjacency matrix was constructed:  $A=[a_{ij}]17 \times 17$ , where  $a_{ij}$  indicates the interaction relationship between the influencing factors. The value of  $a_{ij}$  was 1 when  $S_i$  had a direct influence on factor  $S_j$ ; the value of  $a_{ij}$  was 0 when factor  $S_i$  did not have a direct influence on factor  $S_j$ ; the cells on the diagonal of factor  $S_i$  were all recorded as 0.

b) Determine the recursive relationship between factors.

Step 1: Based on MATLAB platform, final reachable matrix M was calculated.

Step 2: A hierarchical structure division table of the factors influencing the investment estimation of intelligent construction of high-speed railways was sorted out, as shown in Table I.

c) *Construction of an ISM:* The influence factor structure was drawn according to the reachable set, antecedent set, and factor hierarchy decomposition results, as shown in Fig. 1.

3) *Investment estimation module based on case inference and SD:* System dynamics analysis was considered based on

the complex factors and the nonlinear and changeable characteristics of the estimation process.

a) Construction of cause-and-effect diagram of intelligent construction investment estimation system for high-speed railways.

Based on the influence paths obtained from the ISM, the study analyzed the cause-effect relationship, the overall and local feedback mechanism of the system, and the feedback loops in the system for the intelligent construction investment estimation system of high-speed railways and further clarified the polarity of the variables in the system and their mutual influence. The system feedback relationship is shown in Fig. 2.

b) The overall flow chart of the intelligent construction investment estimation system of high-speed railways.

A model flow diagram was established as shown in Fig. 3 by taking the investment estimation cost of intelligent construction of high-speed railways, labor cost, material cost, construction machine use cost, robot cost, enterprise management cost, equipment purchase cost, and the estimated investment cost at the project proposal stage as state variables and taking changes in labor cost, material cost, construction machine use cost, robot cost and enterprise management cost and willingness to invest as rate variables.

4) Trend analysis module for the evolution of investment estimation based on case inference and SD.

On the basis of the above model, the evolution trend of investment estimation was analyzed by inputting 50%, 60%, 70%, 80%, 90%, and 100% intelligence levels.

5) Improvement module for investment estimation derivative mechanism and estimation method.

Based on the above trends of investment estimation evolution, the corresponding derivative mechanisms and laws were deduced, and suggestions for improving labor value theory and the current investment estimation preparation methods were proposed.

TABLE I. RESULTS OF THE HIERARCHICAL DIVISION OF KEY INFLUENCING FACTORS

Levels	Factors
Level 1 (direct apparent influencing factors)	Labor, material and machine costs; management cost
Level 2 (indirect influencing factors)	Project complexity; resource allocation efficiency; professional quality of management staff; management efficiency; degree of economic development in the region; material dynamic management capacity
Level 3 (indirect influencing factors)	Design depth; intelligent robot configuration; construction site management level; traditional construction equipment configuration; information management level
Level 4 (decisive influencing factor)	Intelligent degree of geological exploration; intelligent degree of construction; degree of equipment mechanization; intelligent degree of management

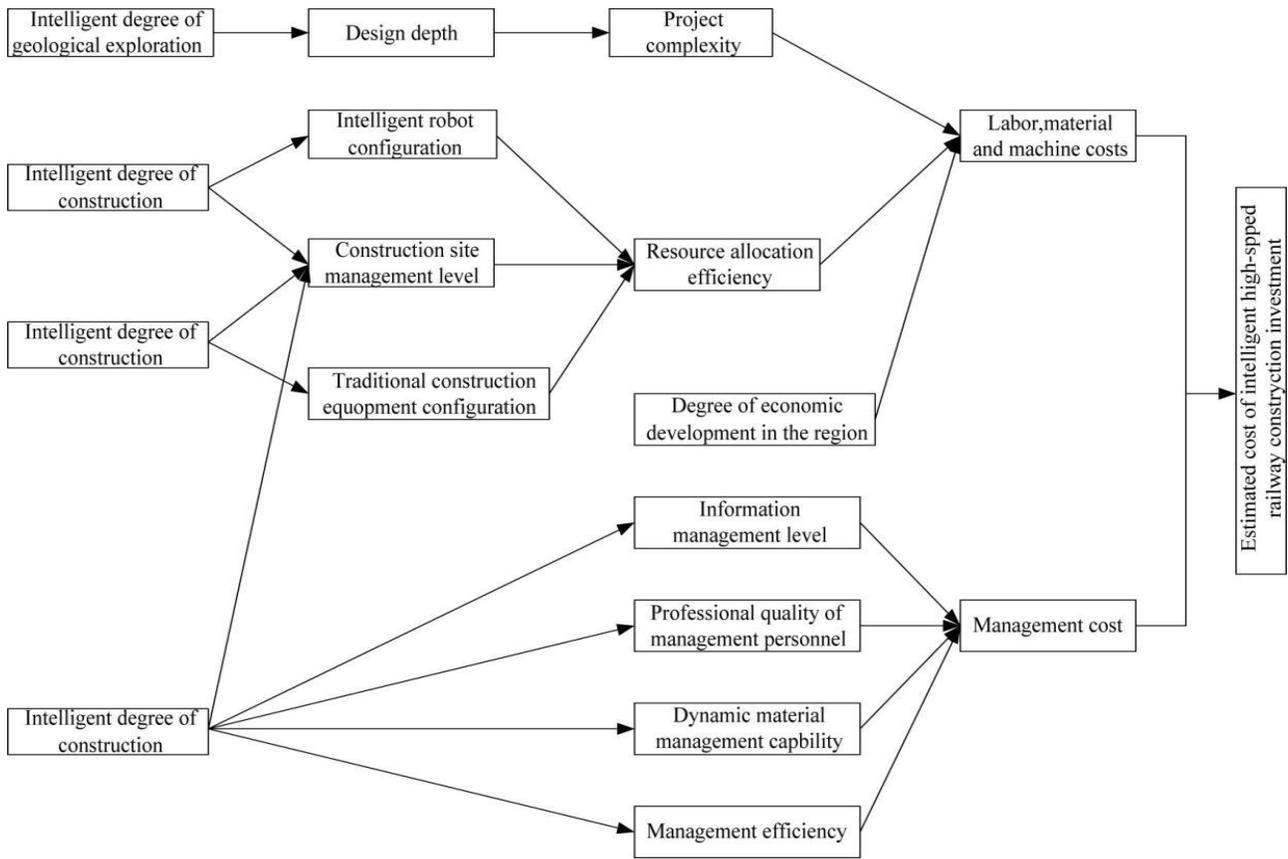


Fig. 1. The ISM of factors influencing investment estimation for intelligent construction of high-speed railways.

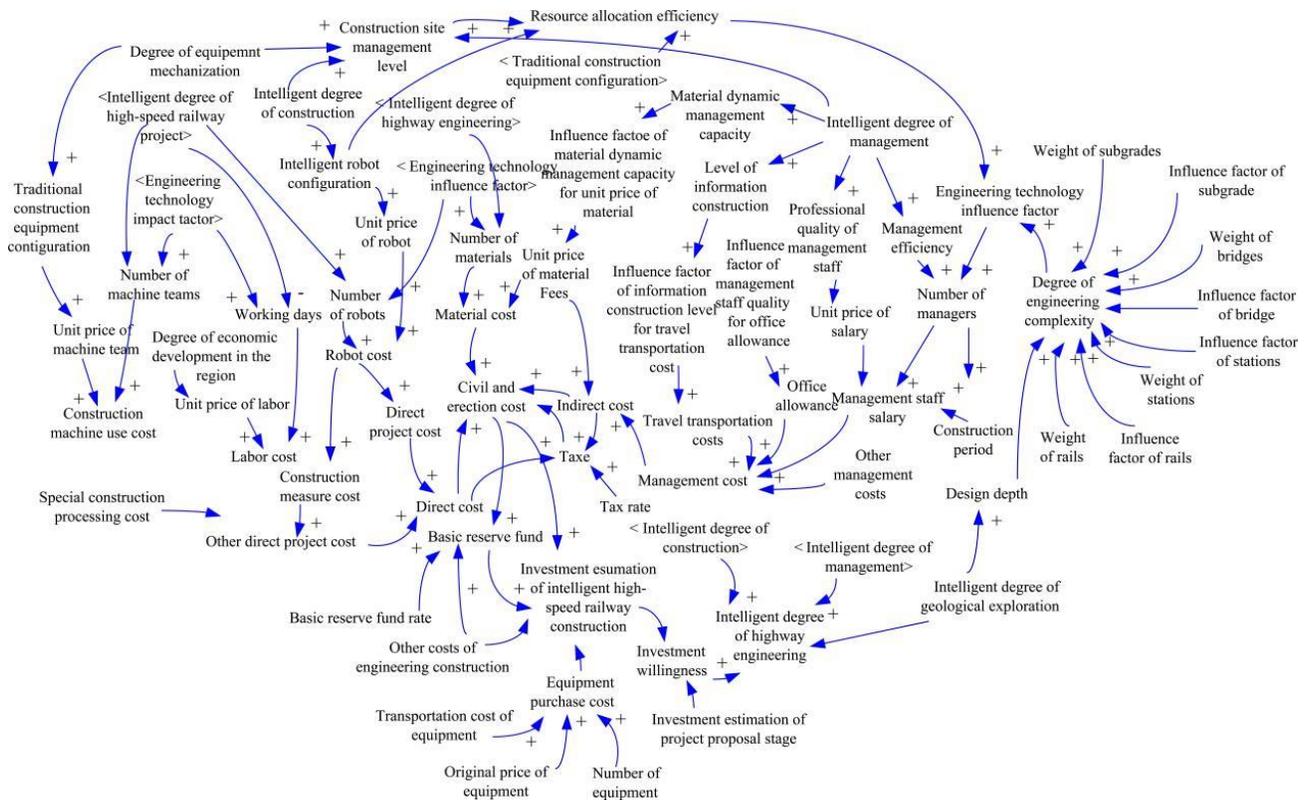


Fig. 2. System feedback relationship diagram.

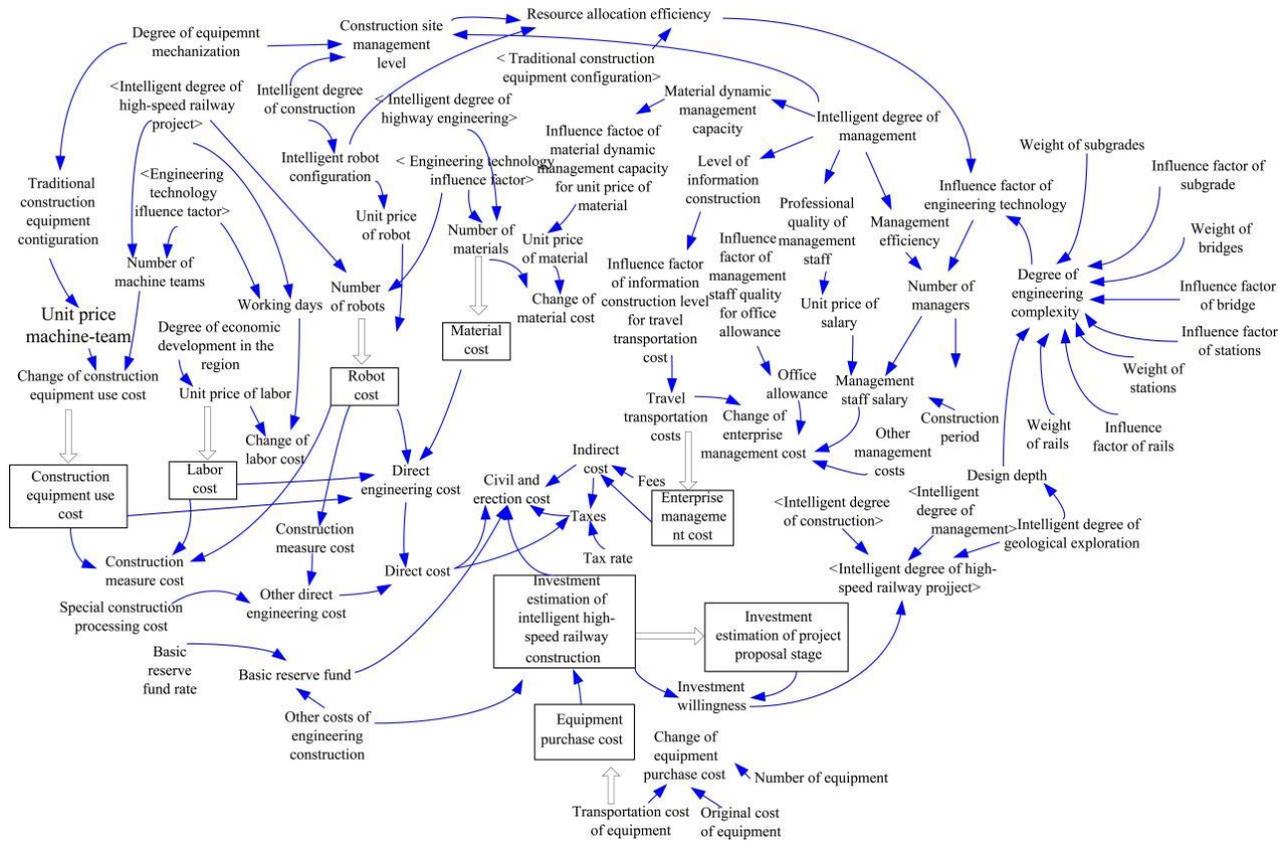


Fig. 3. Model flow diagram.

#### IV. MODEL APPLICATION (INTRODUCTION TO PROJECT COMPLEXITY)

Models were applied in the construction of a high-speed train (HST) project, which started in 2019, with a bridge-tunnel ratio of 94%. It was constructed by mature tunnel intelligent construction technology and is a highly intelligent high-speed railway construction project. System dynamics analysis was applicable due to its complex factors and the non-linear and changeable characteristics of the estimation process. Simulations were performed based on the constructed SD model. The main parameters and equations were set, including constants, initial values, linear functions, and table functions.

##### A. Estimation of Constants and Initial Values

With the extension of the construction period, the costs showed exponential situation growth, and the data involved were poorly available. The above characteristics indicated that it was reasonable to adopt the GM(1,1) model. This paper only shows the prediction of other costs of engineering construction, as shown in Table II, and the remaining costs were predicted in the same way.

The grading ratio test values were all within the standard range interval [0.672,1.501], which met the requirement of the construction of the GM(1,1) model. The results obtained according to equations 4 and 5 are shown in Table III, and test difference ratio  $C = 0.0027 \leq 0.35$ , i.e., the accuracy of the constructed model met the requirement, so other costs spent in engineering construction was predicted using the model, as shown in Table IV.

TABLE II. GRADING RATIO VALUES UNDER THE GM(1,1) MODEL

Case	Original value	Grading ratio $\lambda$	Original value + translation conversion shift value (shift = 0)	Converted grading ratio $\lambda$
1	402.868	-	402.868	-
2	393.157	1.025	393.157	1.025
3	385.965	1.029	385.965	1.029
4	387.421	0.996	387.421	0.996

Note:  $\lambda$  = data in the previous period/data in the current period

TABLE III. THE RESULTS OF GM(1,1) MODEL CONSTRUCTION

Development factor a	Grey action quantity b	Posterior test difference ratio c
-0.0110	377.2902	0.0027

TABLE IV. PREDICTIVE VALUES OF OTHER COSTS OF ENGINEERING CONSTRUCTION

Serial number	Original value	Predicted value
1	402.868	402.868
2	393.157	393.157
3	385.965	385.965
4	387.421	387.421
Backward one period	-	393.123
Backward two periods	-	392.441
Backward three periods	-	410.041
Backward four periods	-	405.263

**B. Estimation of the Values of Factors of the Linear Functional Relationship**

Taking the construction site management level as an example, its subtrees were the degree of management intelligence, the degree of construction intelligence, and the degree of equipment mechanization. The subjective and objective weights were calculated by the comprehensive weighting method by taking these subtrees as the boundary point.

Step 1: determine the subjective weights using the G1 method.

Suppose that  $x_1, x_2,$  and  $x_3$  are the degree of management intelligence, the degree of construction intelligence, and the degree of equipment mechanization, respectively; experts believed that these three elements had the following relationship:  $x_1 \geq x_2 \geq x_3$ , and  $r_2 = \frac{w_1^*}{w_2^*} = 1.2, r_3 = \frac{w_2^*}{w_3^*} = 1.2, r_2 \cdot r_3 = 1.44, r_3 = 1.2, \sum_{k=2}^3 (\prod_{i=k}^3 r_i) = 2.64$ .

According to Equations 7, 8, and 9,  $w_1 = 0.33, w_2 = 0.396,$  and  $w_3 = 0.275$ .

Step 2: Determine objective weights by the entropy method.

The scores given by experts are shown in Table V.

TABLE V. SCORES GIVEN BY EXPERTS

Factors	Expert 1	Expert 2	Expert 3	Expert 4	Expert 5
Degree of management intelligence	0.2	0.2	0.1	0.2	0.2
Degree of construction intelligence	0.3	0.2	0.3	0.3	0.2
Degree of equipment mechanization	0.4	0.4	0.3	0.3	0.4

The calculation results of equation 10 are shown in Table VI.

TABLE VI. FEATURE WEIGHT

$j$	$p_{ij}$					$e_j$	$g_j$	$w_j$
1	0.222	0.222	0.111	0.222	0.111	1.581	0.581	0.327
2	0.231	0.154	0.231	0.231	0.154	1.591	0.591	0.334
3	0.222	0.222	0.167	0.167	0.222	1.6	0.6	0.338

Step 3: determine the final weights using the comprehensive weighting method.

According to equations (11) and (12),  $k_1 = 0.669, k_2 = 0.715$ . The final weights were calculated using equation (13). After normalization,  $w_1 = 0.328, w_2 = 0.365,$  and  $w_3 = 0.307$ , i.e., construction site management level =  $0.328 * \text{the degree of management intelligence} + 0.365 * \text{the degree of construction intelligence} + 0.307 * \text{the degree of equipment mechanization}$ .

**C. Estimation of the Values of Factors of the Table Function Relationship**

It was assumed that the data on the degree of management intelligence were accurate; the dynamic management capability of materials would be improved as the degree of management intelligence increased. The corresponding table function was established according to the statistical law, as shown in Table VII.

TABLE VII. TABLE FUNCTION RELATIONSHIP

Management intelligence degree	0	0.1	0.2	0.3
Material dynamic management capability	0.73	0.78	0.83	0.89

The material dynamic management capability = WITH LOOKUP (the degree of management intelligence) Lookup([(0,0)-(1,1)], (0,0.73), (0.1,0.78), (0.2,0.83), (0.3,0.89)).

Resource allocation efficiency and management efficiency were represented by linear functions and table functions. The model constructed in this paper successfully passed the structure, assignment, and unit consistency tests.

**D. Model Simulation and Result Analysis**

Simulation settings were performed in Vensim software to simulate the intelligent construction investment estimation results of the XC section of the HST high-speed railway. The comparison between the results and the actual data of the project suggested that the error between the estimated cost of the intelligent construction investment of high-speed railway in the first year simulated by the system and the actual value was 4.62%; the error between the two-year cumulative investment estimation and the actual investment was 1.83%. This suggested that the accuracy met the requirements, and the model simulation results could be considered to be in line with the reality.

The experimental results of the SD model proposed in this paper were compared with the experimental results of other high-speed railway intelligent construction investment estimation models that can be found in the related literature, i.e., the grey-wolf optimizer-support vector machine (GWO-SVM) estimation model [19] and the improved back-propagation neural network (BPNN) prediction model [20], and the specific data are shown in Table VIII. The error rate in Table VIII is the error between the two-year cumulative investment estimation and the actual investment, and it was seen that the error rate of this model was 1.83%, which was smaller other estimation models. It verified the accuracy of the SD model in the field of investment estimation for intelligent construction of high-speed railway.

TABLE VIII. THE EXPERIMENTAL RESULTS OF DIFFERENT ESTIMATION MODELS

Model	Error rate
The SD model	1.83%
GWO-SVM estimation model	3.14%
The improved BPNN prediction model	3.23%

V. SCENARIO SIMULATION OF THE EVOLUTIONARY TREND OF THE ESTIMATED COST OF HIGH-SPEED RAILWAY INVESTMENT UNDER HIGH INTELLIGENCE

The investment cost was simulated by taking the degree of management intelligence as an example to review the composition of the investment estimation content of the intelligent construction of high-speed railways. Intelligent management refers to the management of the whole process of construction with the help of information management system in the process of high-speed railway construction management. As the degree of intelligent management increases, fewer managers are required, and corporate overheads are reduced. Comprehensive intelligence degree of the high-speed railway project = IF THEN ELSE (investment willingness  $\leq$  0.45,  $0.65 \times$  the intelligent degree of geological exploration +  $0.2 \times$  the intelligent degree of construction +  $0.15 \times$  the intelligent degree of management,  $0.3 \times$  the intelligent degree of geological exploration +  $0.35 \times$  the intelligent degree of construction +  $0.35 \times$  the intelligent degree of management).

The relationship between investment costs and the change of comprehensive intelligence degree of high-speed railway project is shown in Table IX. The labor cost and management cost decreased with the improvement of the intelligence degree, while the robot cost increased significantly, as shown in Fig. 4 and 5. This deviated from the content of traditional railroad investment estimation, so the estimation content and taxable base need to be improved.

TABLE IX. VARIATIONS OF DIFFERENT COSTS WITH THE IMPROVEMENT OF THE COMPREHENSIVE INTELLIGENCE OF HIGH-SPEED RAILWAY PROJECT

Degree of intelligence	Labor cost	Material cost	Construction machine use cost	Robotic cost	Management cost	Profit	Tax
50%	2210.45	15746.13	2746.83	2.29	63.07	1453.81	2444.48
60%	1822.25	14849.78	2471.82	13.13	50.44	1344.52	2260.71
70%	1515.99	14216.74	2236.55	18.71	40.15	1261.97	2121.91
80%	1145.99	13036.42	1850.11	25.68	31.10	1126.25	1893.71
90%	457.54	12065.18	1554.47	224.97	20.57	1002.59	1685.78
100%	0	11406.18	1065.27	492.93	9.70	908.19	1527.05

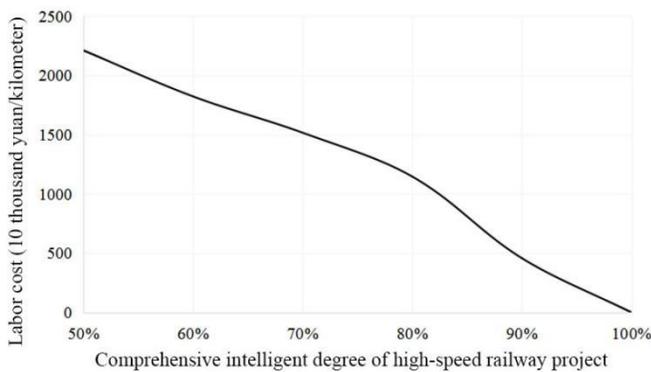


Fig. 4. Trend graph of labor cost change.

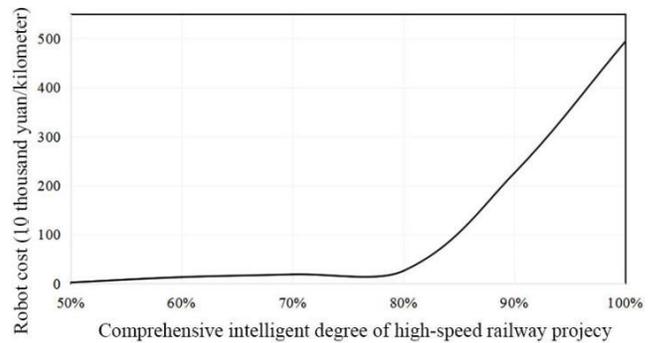


Fig. 5. Trend graph of robot cost change.

VI. STUDY ON THE DERIVATION MECHANISM OF HIGH-SPEED RAILWAY INVESTMENT ESTIMATION UNDER HIGH INTELLIGENCE

At present, the theoretical basis of high-speed railway intelligent construction investment estimation is labor value theory, and machines and robots are only materialized labor, which is only transfer value; the calculation method theoretically takes the increased value of labor cost, management cost, and depreciation cost as the profit and tax billing base, but actually the profit and tax are calculated based on the cost excluding tax. Therefore, under 70-100% high intelligence, the estimated cost and calculation method of intelligent construction investment of high-speed railways need to be revised. Based on this, this paper proposed the amendment of intelligent civil and erection costs and the amendment of tax rate increase.

A. Intelligent Civil and Erection Cost Content Amendment

The cost components and content of intelligent civil and erection costs under high intelligence were adjusted using value theory, system theory and artificial intelligence. The specific elaboration is shown below.

- 1) *Material cost*: With the deepening of construction site intelligence, the dynamic management ability of materials is improved, and thus the material cost is reduced.
- 2) *Construction machine use cost*: Traditional construction machines are replaced by intelligent equipment such as robots on a large scale to significantly reduce construction machine use cost.
- 3) *Robot cost*: A new item of “robot cost” was created, mainly for robot depreciation, while construction equipment use cost becomes the cost of using traditional equipment.
- 4) *Enterprise management cost*: With the continuous improvement of the degree of intelligence, enterprise management cost is greatly reduced; when complete intelligent is realized, enterprise management cost will include the senior management salary.
- 5) *Profits and taxes*: In the case of complete intelligence, there are few labor and management costs. If the robots are still regarded the transfer value of physical labor and excluded from the VAT base, it will seriously affect the profits of enterprises and VAT, resulting in a significant reduction in profits and value added of enterprises.

### B. Rate Increase Amendment

With the development of intelligent production, robots, like people, are a source of surplus value creation. If the rate is not changed, it will seriously affect the profits of enterprises and taxes such as VAT, which will affect the national tax revenue. Therefore, the tax rate should be increased, and profits and VAT should be calculated according to the current cost excluding tax.

### C. Improvements and Supplements to the Labor Theory of Value

1) *Analysis of unmanned construction value sources:* The overall value ( $W$ ) of a conventional railroad investment should be expressed as:  $W = C + V + M$ , where  $C$  denotes constant capital,  $V$  denotes variable capital, and  $M$  is surplus value. Due to intelligence and unmanned construction, the robot's production is a mature production chain. In the production process, the means of production ( $C'$ ) and the mental and physical labor of scientists and workers ( $V'$ ) are all necessary and are also the source of surplus value ( $M'$ ), and at this moment the value of the robot is:  $W' = C' + V' + M'$ . Then, the value generated by the intelligent high-speed railway can be written as:  $W = W' + C_2 = C' + V' + M' + C_2$ . Based on this, when the intelligent construction of the high-speed railway is carried out, the cost of purchasing robots already includes the mental labor of the scientists and the physical labor of the workers who produce these robots.

2) *Changes to the labor value theory brought about by the fully intelligent construction of high-speed railways:* The only source of surplus value is still labor, but the source of labor has changed to labor and robot capital, which is the inevitable result of the rapid development of intelligence. With the development of intelligent technology, the demand for labor is getting lower and lower, and it may even produce the phenomenon of creating robots through the labor of robots in the future. Nevertheless, it is undeniable that although human labor creation is no longer reflected in every production line, the source of surplus value creation is still human, and human and human labor must be the source of value creation, which is an indispensable condition for the creation of artificial intelligence robots.

## VII. CONCLUSION

This paper briefly introduced the ISM and the SD model. The relationship between the factors affecting the intelligent construction investment estimation of high-speed railway was sorted through the ISM model, and the SD prediction and improvement model was constructed for the intelligent construction investment estimation of high-speed railway. Simulation experiments were carried out in the construction of HST high-speed railway. It was found that the error between the first-year estimated cost of intelligent construction investment in system simulation and the actual value was

4.62%; the error between the two-year cumulative investment estimate and the actual investment was 1.83%. The results prove that the SD model has high accuracy and is feasible in the investment estimation of intelligent construction of high-speed railway.

### ACKNOWLEDGMENTS

This study was supported by Hebei Natural Science Foundation: Research on 3D Nonlinear Intelligent Estimation Method of Investment and Operation Income of High Speed Railway (G2019210226).

### REFERENCES

- [1] S. M. Tian, W. Wang, C. Y. Yang, Y. Liu, N. Y. Wang, K. J. Wang, Z. F. Ma, and G. Lu, "40 year development and Prospect of railway tunnels in China," *Tunn. Constr.*, vol. 41, pp. 1903-1930, 2021.
- [2] K. F. Wang, R. Hao, W. L. Lu, and H. L. Wang, "Research and application of intelligent construction technology in railway engineering construction," *China Railw.*, pp. 45-50, 2019.
- [3] Deutsche Bahn-Digital Transformation and Long-term Challenges. 2018.
- [4] P. Li, S. Shao, R. Xue, and X. D. Zhang, "Research on the development trend of digitalization and intellectualization of foreign railways," *China Railw.*, pp. 25-31, 2019.
- [5] European Commission. Road map to a Single European Transport Area Towards a competitive and resource efficient transport system. (2011-03-28) [2018-10-24].
- [6] The European Rail Research Advisory Council. Rail route 2050: The Sustainable Backbone of the Single European Transport Area. (2011) [2018-10-24].
- [7] European Commission. Horizon 2020 Work Programme 2016—2017 Smart, green and integrated transport. EU, 2015.
- [8] East Japan Railway Company. The Mid-to-Long-term Vision for Technological Innovation. (2016) [2018-11-06].
- [9] L. Zhou, "Construction of engineering cost model based on system dynamics," *Build. Technol. Dev.*, vol. 48, pp. 121-123, 2021.
- [10] X. Q. Mao, Wei. Li, and C. H. Mao, "System dynamics analysis on the promotion path of multi-agent synergy effect of engineering construction under BIM," *J. Civ. Eng. Manag.*, vol. 37, pp. 80-85, 2020.
- [11] S. L. Yue, and W. Zhang, "Research on construction project quality management based on system dynamics," *Sci. Technol. Econ. Market*, pp. 97-98, 2020.
- [12] J. Gao, and X. T. Wu, "Risk evolution evaluation of engineering projects based on system dynamics model," *Stat. Decis. Making*, vol. 36, pp. 185-188, 2020.
- [13] Y. Zhong, Z. G. Chen, and Z. Zhou, "Research on resource allocation model and strategy of large-scale construction projects -- modeling and simulation based on system dynamics," *China Manage. Sci.*, vol. 24, pp. 125-132, 2016.
- [14] J. K. Liu, and W. J. Yang, "Research on cost schedule control of construction projects based on system dynamics," *Constr. Technol.*, vol. 45, pp. 95-99 + 128, 2016.
- [15] Y. Chen, Y. Wang, and L. Jia, "Research on cost control of prefabricated buildings based on system dynamics," *Value Eng.*, vol. 36, pp. 1-5, 2017.
- [16] A. Akintoye, "Analysis of factors influencing project cost estimating practice," *Constr. Manag. Econ.*, vol. 18, pp. 77-89, 2000.
- [17] W. Duan, Y. Qi, F. Gong, and D. S. Xu, "A review of research on the combination of system dynamics and economic management theories and methods," *Stat. Decis. Making*, vol. 38, pp. 41-46, 2022.
- [18] B. Jia, "Research on risk management of SH key engineering projects based on interpretive structural model," Jinan: Shandong University, 2020.

- [19] X. Chen, "Investment probabilistic interval estimation for construction project using the hybrid model of SVR and GWO," *J. Constr. Eng. M.*, vol. 147, pp. 1-13, 2021.
- [20] B. Wang, "Discussion on the prediction of engineering cost based on improved BP neural network algorithm," *J. Intell. Fuzzy Syst.*, vol. 37, pp. 6091-6098, 2019.

# Early Warning for Sugarcane Growth using Phenology-Based Remote Sensing by Region

Sudianto Sudianto<sup>1</sup>, Yeni Herdiyeni<sup>2</sup>, Lilik Budi Prasetyo<sup>3</sup>

Department of Informatics, Institut Teknologi Telkom Purwokerto, Indonesia<sup>1</sup>

Department of Computer Sciences, IPB University, Indonesia<sup>2</sup>

Department of Forest Resources Conservation & Ecotourism, IPB University, Indonesia<sup>3</sup>

**Abstract**—It is crucial to know crop growing in order to increase agricultural productivity. In sugarcane's case, monitoring growth can be supported by remote sensing. This research aimed to develop an early warning for sugarcane growth using remote sensing with Landsat 8 satellite at a crucial phenological time. The early warning was developed by identifying regional sugarcane growth patterns by analyzing seasonal trends using linear and harmonic regression models. Identification of growth patterns aims to determine the crucial phenological time by calculating the statistical value of the NDVI spectral index. Finally, monitoring the sugarcane growth conditions with various spectral indices for verification: NDVI, NDBaI, NDWI, and NDDI. All processes used Google Earth Engine (GEE) as a cloud-based platform. The results showed that sugarcane phenology from January to June is crucial for monitoring and assessment. The value of the four corresponding indices indicated the importance of monitoring conditions to ensure a healthy sugarcane region. The results showed that two of the four regions were unhealthy during particular periods; unhealthy vegetation values were below 0.489 and vice versa, one due to excess water and the other due to drought.

**Keywords**—Google earth engine; landsat 8; monitoring and assessment; sugarcane health

## I. INTRODUCTION

Sugarcane is a raw material for the sugar industry that plays a strategic role in Indonesia's economy, with a crop area of 413.05 thousand ha in 2019 [1]. Sugar consumption in Indonesia continues to increase. In 2023, it will reach 7.15 million tons, in line with Indonesia's population growth rate [2]. However, domestic sugar production has not been able to meet Indonesia's sugar needs [2]. In response to the increasing demand for sugar, Indonesia has declared the aim of self-sufficiency by increasing sugarcane production and productivity through policies and programs in the "on-farm" aspect [3]. In an effort to support the sugarcane production and productivity in Indonesia, it is essential to apply the proper technology, including remote sensing by satellite for early warning of sugarcane growing at crucial phenological times.

An early warning detects potential incidents and is beneficial in managing and preventing factors that cause crop loss, such as insufficient or excessive irrigation, disease, and pests. In addition, early warning at crucial phenological times helps identify and measure factors that affect plant growth, especially sugarcane. Through early warning, sugarcane regions experiencing stress can be identified and classified so that farmers can change their practices to increase sugarcane

productivity and production. Then, knowing the condition of plants at the peak of phenology can also help in agricultural planning and management.

The application of remote sensing to sugarcane crop conditions has evolved in the last decade, for example, monitoring the growth phase of sugarcane by satellite in West Java, Indonesia [4]; the role of NDVI in mapping sugarcane conditions around oil and gas fields [5]; assessment of sugarcane conditions based on NDVI [6]; monitoring sugarcane growth [7]; sugarcane yield estimation and forecasting in smallholder farming conditions [8]; prediction of sugarcane yield based on NDVI and nutrient concentration [9]; and prediction of crop yields from MODIS relative vegetation health in Africa [10]. However, to monitor sugarcane growth, there has been no early warning during the crucial times of sugarcane phenology. In previous studies, plant growth was only through a cooperative data approach, not based on plant phenology knowledge. In addition, the method in the previous study was based on vegetation parameters only. Meanwhile, in this study, plant growth with several parameters based on satellite data was built with linear regression based on phenological knowledge.

On the other hand, available and open remote sensing data can now be analyzed with more affordable computing, and a free cloud platform can be used for geospatial analysis, namely Google Earth Engine (GEE). GEE, as a cloud platform, is efficient for geospatial analysis [11], which is helpful for precision agriculture, with its availability of comprehensive and open data, for example, for monitoring vegetation's current state and dynamics [12]. GEE applications have also been widely used to solve agricultural problems, for instance, mapping sugarcane by integrating multitemporal Sentinel-2 images [13]; rice mapping based on SNIC segmentation [14]; and object-oriented crop classification [15].

This research offers a solution to issues of production and productivity in building an early warning for sugarcane conditions at crucial phenological times through monitoring and assessment. The method of creating an early warning include several approaches: (1) analyzing seasonal trends in the region using linear regression and harmonic models to identify sugarcane growth patterns; (2) assessing the condition of sugarcane through statistical analysis of the Normalized Difference Vegetation Index (NDVI) spectral index at the crucial phenological phase; and (3) monitoring sugarcane conditions with various spectral indices, based on typical use, and having accurate results for condition interpretation, with

NDVI as the vegetation index, Normalized Difference Bareness Index (NDBaI) as the fallow land identification, Normalized Difference Water Index (NDWI) as a water content indicator, and Normalized Difference Drought Index (NDDI) as a drought indicator. The four spectral indices serve as a system of checks and balances to assess sugarcane conditions. All processes were performed on GEE as a cloud-based platform. Therefore, this research aimed to develop an early warning of sugarcane growth using remote sensing at a crucial phenological time. From the solutions offered, farmers can better anticipate particular practices, and the government can increase productivity and sugarcane production in a more sustainable manner through precision farming.

This paper is described in systematics: materials and methods used for early warning with several parameter indicators, building knowledge of plant phenology; and the results section to determine plant status using remote sensing data. The conclusions and recommendations section reports findings for future research.

## II. MATERIALS AND METHODS

### A. Study Area

This study was carried out at sugarcane plantations of Djengkol Kediri, East Java Province, Indonesia, with a coordinate polygon of (Lat: 112.199, Long: -7.892). The area was selected because (1) East Java is the largest producer in Indonesia, with Djengkol Kediri as the second largest [2]; (2) the plantations in the area support the economy, and are the basis for the sugar self-sufficiency program; (3) it is a representation and model for other fields. As shown in Fig. 1.

In the study area (Fig. 1), the region was selected based on the sugarcane plantation and divided into four regions: (1) region A: relatively flat terrain with a mean elevation of 260 m above sea level; (2) region B: relatively flat terrain with a height of 264 m; (3) region C: sloping terrain (left to right) with a mean elevation of 267–272 m; (4) region D: sloping terrain (left to right) with a mean elevation of 273–276 m. The four regions are planted with the same sugarcane variety.

### B. Maintaining the Integrity of the Specifications

The image data used in this study were sourced from the Landsat 8 satellite, available at GEE “LANDSAT/LC08/C01/T1\_TOA”. The public image collections were filtered with by using several timescales and criteria: (1) image collection from 2015 to 2020 was used to analyze seasonal sugarcane trends to obtain cropping patterns in the study region and to mask clouds; (2) cloud-free image data from 2014 to 2017 were used for the assessment of sugarcane conditions; (3) 2019–2020 was used for monitoring data on sugarcane health and cloud-free image collection. In addition, all image collection timescales used a resolution of 15 m, increasing the resolution shown in Fig. 2 through the Brovey transformation approach by blending the panchromatic [16] (Formula 1):



Fig. 1. The study area in the sugarcane plantation of Djengkol Kediri, Indonesia

$$R_{in} = R_{in} / (R_{in} + G_{in} + B_{in}) \times P_{in} \quad (1)$$

where  $R_{in}$  = Band 4,  $G_{in}$  = Band 3,  $B_{in}$  = Band 2, and  $P_{in}$  = Band 8 (panchromatic).



Fig. 2. RGB image 30 m (left) and the result of image sharpening through panchromatic blending, 15 m (right)

### C. Sugarcane Phenology

Plant phenology can identify and obtain important information about vegetation [17]–[19]. Sugarcane (*Saccharum officinarum*) is a type of grass with unique characteristics. Sugarcane phenology has four phases: (1) germination—this phase lasts for 15–20 days while new shoots are growing; (2) tillering—the sugarcane plants grow tillers for 4–6 months during this phase, and up to 50% of each sugarcane stalk grows leaves; (3) grand growth during this phase, lasting for 5 months, sugarcane height elongates and plants grow to maturity; and (4) maturity: this phase, which occurs during the 3 months before harvest, includes vegetative decline and sucrose accumulation of up to 55% per dry weight of sugarcane [20]. The leaf area index (LAI) value reaches its maximum when the plant is 6 months old, decreasing slowly afterward [6], [20].

Sugarcane phenology generally lasts 10–12 months. In this study area, sugarcane phenology (Table I) proceeds as follows: from early October to November is seeding and budding, January to April is stem elongation, May to August is the period of sugar accumulation and ripening, and September is the harvest period.

TABLE I. TIME OF SUGARCANE PHENOLOGY IN STUDY AREA

Crop	Years/Month	Periods	
Sugarcane	2019	Oct	SE
		Nov	SE
		Dec	SL
	2020	Jan	ST
		Feb	ST
		Mar	ST
		Apr	ST
		May	SA
		Jun	MA
		Jul	MA
		Aug	MA
Sep	HA		

Note: HA: harvest stage; SE: seeding stage; SL: seedling stage; ST: stem elongation stage; SA: sugar accumulation stage; MA: maturation stage.

D. Spectral Indices

Plants can absorb and reflect unique light waves. This phenomenon is carried out by chlorophyll in the mesophyll tissue of leaves (photosynthesis) [21]. As a basis for remote sensing using satellites, the brightness value received by satellite sensors in a particular band was used to identify plants. The spectral index works by calculating the wavelength of the band composition.

There are many types of indices, and only four were used in this study, namely (1) Normalized Difference Vegetation Index (NDVI), which describes the greenness of a plant. It is based

on a mathematical combination of visible red light and near-infrared radiation (NIR) channels, which are used as indicators of the presence and condition of vegetation [22], as illustrated in Formula 2; (2) Normalized Difference Bareness Index (NDBaI) is used to separate fallow or open land from other types of cover using Landsat image data [23]. It is very sensitive for distinguishing between fallow, semi-fallow, and cultivated areas. The index uses the short infrared (SWIR) and thermal infrared (TIR) bands, as shown in Formula 3; (3) Normalized Difference Water Index (NDWI) is used to determine the water content of vegetation [24]. It is obtained from a combination of green and red bands, as presented in Formula 4; (4) Normalized Difference Drought Index (NDDI) aims to identify the presence of drought in vegetation [25] based on a comparison of the NDVI and NDWI index values, as illustrated in Formula 5.

$$NDVI = (NIR - RED) / (NIR + RED) \tag{2}$$

$$NDBaI = (SWIR - TIR) / (SWIR + TIR) \tag{3}$$

$$NDWI = (NIR - SWIR) / (NIR + SWIR) \tag{4}$$

$$NDDI = (NDVI - NDWI) / (NDVI + NDWI) \tag{5}$$

where NIR = Band 5, RED = Band 4, SWIR = Band B6, and TIR = Band 10.

The NDVI spectral index value was used as the primary reference for monitoring sugarcane conditions obtained from the assessment results, while the other three spectral indices (NDBaI, NDWI, and NDDI) were used for verification.

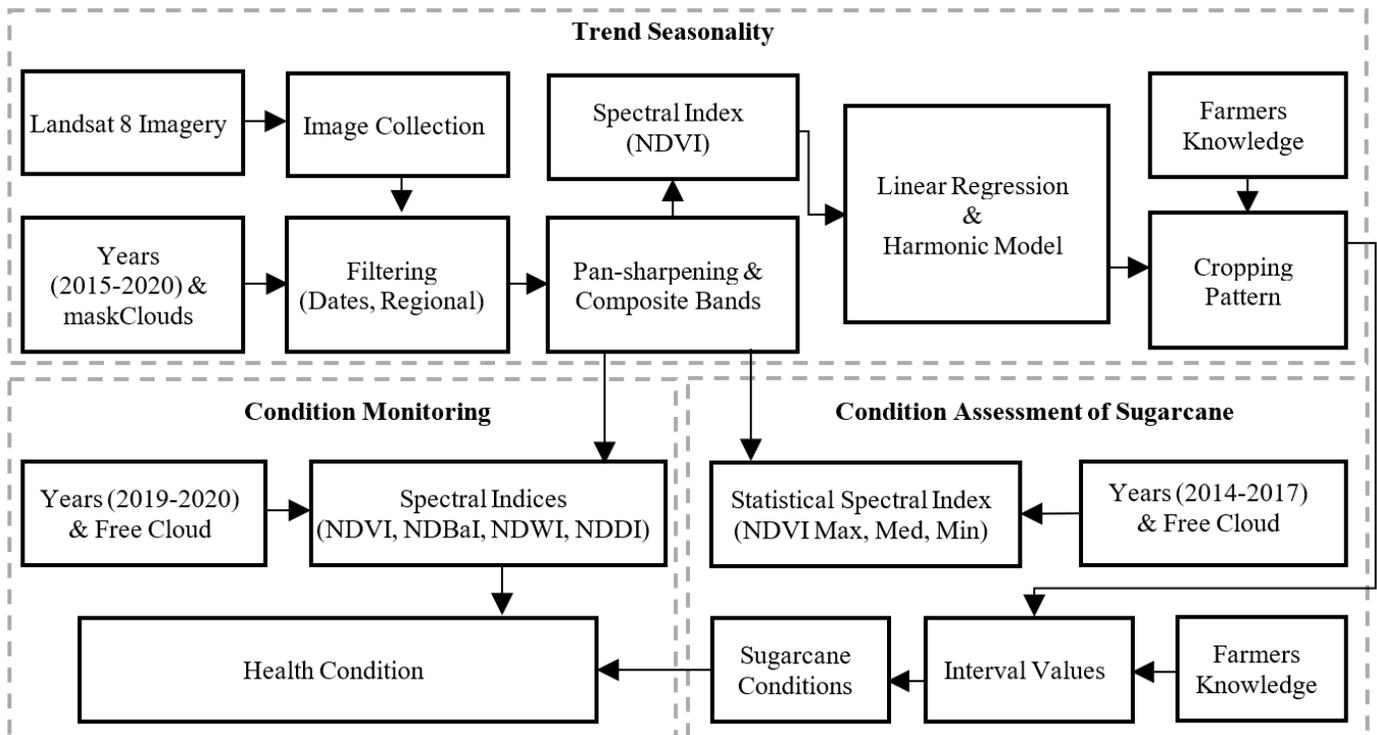


Fig. 3. The study area in the sugarcane plantation of Djengkol Kediri, Indonesia

E. Methodology

Three primary processes were carried out to create an early warning for sugarcane planting conditions (Fig. 3): (1) identification of sugarcane growing patterns by analyzing seasonal trends using a harmonic model (NDVI input); (2)

assessing the condition of sugarcane using the NDVI spectral index; and (3) monitoring sugarcane condition with various spectral indices (NDVI, NDBaI, NDWI, and NDDI). All processes were performed on GEE as a cloud-based platform.

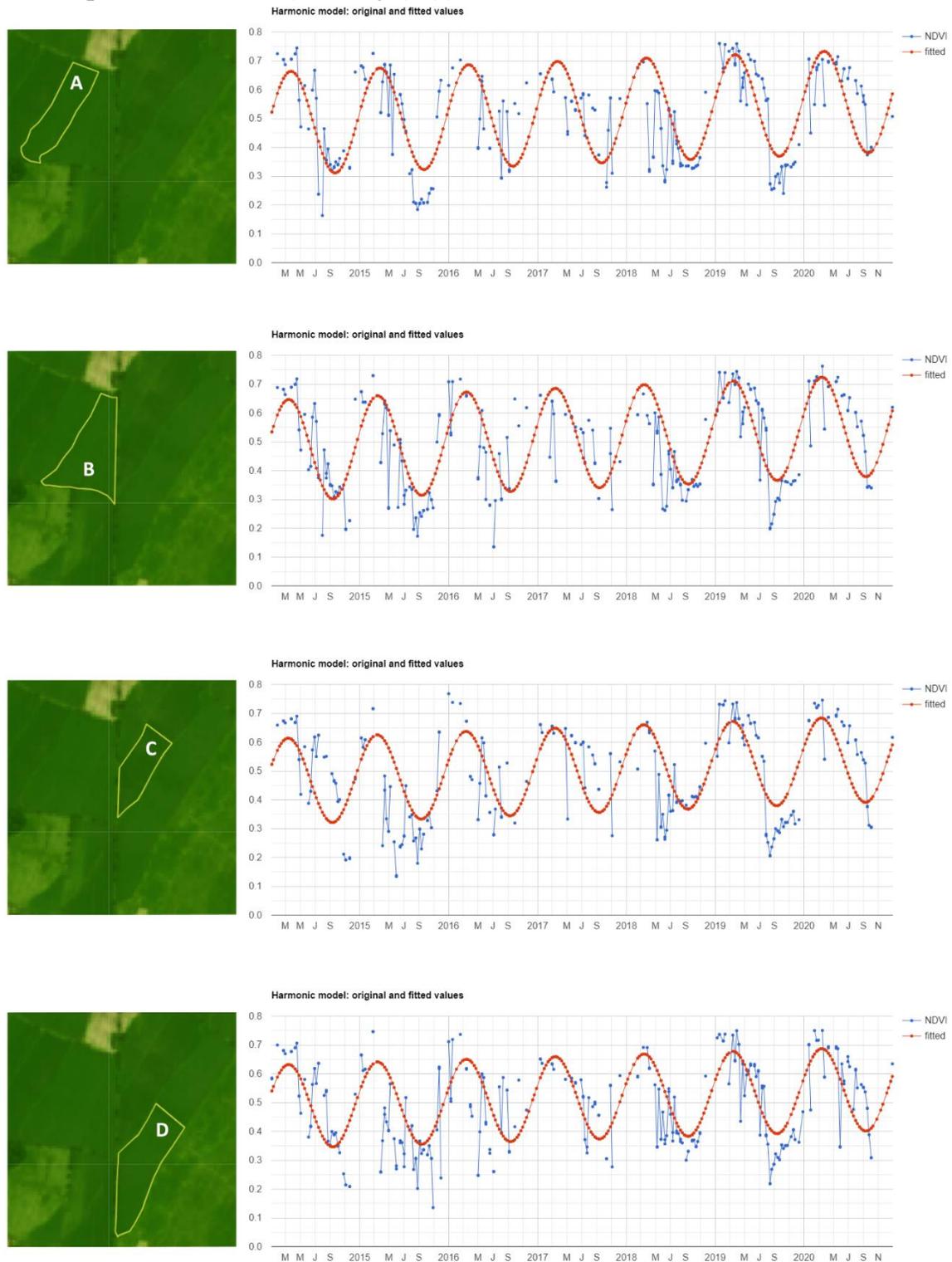


Fig. 4. The frequency of the sugarcane planting pattern based on NDVI values

F. Trend Seasonality

Monitoring seasonal changes in vegetation activity and plant phenology is very important in condition monitoring. In sugarcane, it is necessary to analyze seasonal trends based on phenology using multitemporal data. In GEE, a seasonal trend analysis can be constructed with time series data through temporal data (NDVI) with linear regression (Formula 6) [26]. The seasonal forecast is then built by combining the linear model with the harmonic model (Formula 7) [26].

NDVI values were integrated and used as the basis for seasonal trend analysis (Fig. 4) from image collection data for 2015–2020 in each region (A, B, C, D) by filtering out the masking clouds.

$$p_t = \beta_0 + \beta_1 t + e_t \tag{6}$$

$$p_t = \beta_0 + \beta_1 t + A \cos(2\pi\omega t - \varphi) + e_t$$

$$= \beta_0 + \beta_1 t + \beta_2 \cos(2\pi\omega t) + \beta_3 \sin(2\pi\omega t) + e_t \tag{7}$$

where  $A$  is the amplitude,  $\omega$  is the frequency,  $e_t$  is the random error, and  $\varphi$  is the phase.  $\beta_2 = A \cos(\varphi)$ , and  $\beta_3 = A \sin(\varphi)$ , implying  $A = (\beta_2^2 + \beta_3^2)^{1/2}$ , and  $\varphi = \tan^{-1}(\beta_3/\beta_2)$ . To fit this model to the time series, we set  $\omega = 1$  (one cycle per unit time) and use ordinary least squares regression.

G. Sugarcane Assessment

The purpose of the assessment was to determine the condition of sugarcane. This was based on crucial phenology and used as a reference for an early warning of sugarcane conditions that could predict possible problems emerging in the sugarcane crop. This assessment was developed on phenological knowledge obtained from planting patterns (from analysis of seasonal trends). The results showed the importance of assessing the condition of sugarcane at intervals from January to June. There were several procedures included to assess sugarcane conditions: (1) image data were used for assessment of sugarcane conditions based on cloud-free image

collection data from 2014 to 2017; (2) the locations used were the four regions (A, B, C, D); (3) NDVI spectral information was used as a parameter to determine the condition of sugarcane, then statistical calculations were performed (maximum, median, minimum); (4) the NDVI spectral value of the Landsat 8 image was verified with the farmers’ knowledge on the current condition within the region (A, B, C, D) under prior normal conditions.

III. RESULTS AND DISCUSSIONS

A. Sugarcane Cropping Pattern

Information on sugarcane cultivation patterns in the study area was obtained from the analysis of seasonal trends. The sugarcane cropping pattern was identified through the NDVI value by harmonic model analysis (Fig. 4). The planting pattern became the basis of knowledge for the monitoring and assessment of sugarcane. The sugarcane planting pattern in the study area follows: (1) the cropping pattern begins in October and ends in September of the following year, which applies to all regions (A, B, C, D), as verified with local farmers’ knowledge of appropriate conditions (Table I); (2) the peak NDVI index value occurs six months after planting; (3) the crucial phenology occurs from January to June; therefore, this period was used as a reference for monitoring sugarcane conditions (from two months before peak plant growth to three months after, in preparation for harvest).

B. Assessment of Sugarcane Condition

Sugarcane conditions were assessed from the values of the NDVI spectral index. Based on statistical calculations of the NDVI value of the entire region (A, B, C, D) and the results of local knowledge of the planting pattern, the condition of sugarcane crops from January to June was the focus of the early warning. The interval value for each of those months was the reference for monitoring sugarcane conditions, as shown in Fig. 5.

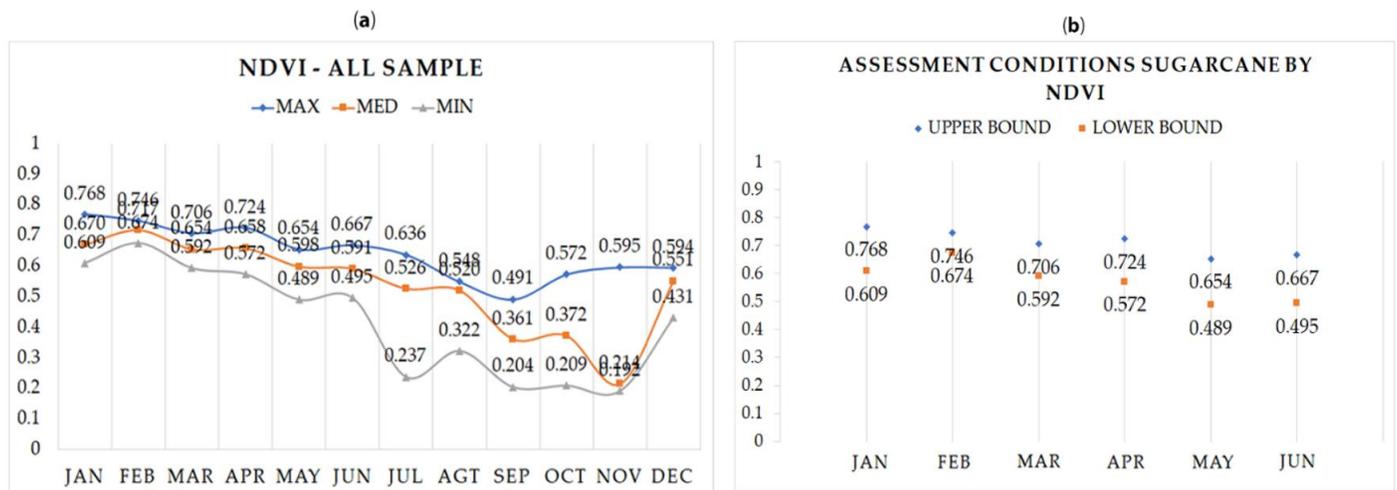


Fig. 5. Results of sugarcane condition assessment: (a) The statistical values from the combination of all sample regions; (b) Sugarcane conditions based on statistical calculations of NDVI values in the crucial months of January to June

C. Assessment of Sugarcane Health

The sugarcane plants were monitored from January to June in 2019–2020 (two periods of planting). The sugarcane condition was measured based on the NDVI spectral index value generated from the sugarcane condition assessment process. Furthermore, to verify the condition, comparisons were made through other spectral indices such as NDBaI, NDWI, and NDDI. The condition of healthy vegetation is the ability to absorb blue- and red-light energy to trigger photosynthesis and create chlorophyll.

Plants with more chlorophyll can reflect more near-infrared energy than the unhealthy variants [21], [27], hence, the spectra of absorption and reflection in visible. Thus, plant spectra of absorption and reflection in visible and infrared wavelengths can provide information about plant health and productivity.

From the results of monitoring the condition of sugarcane plants (Fig. 6), and verified by other spectral values (NDBaI, NDWI, and NDDI), sugarcane condition monitoring was categorized as follows:

- Regions A and B were in good health. The NDVI value belonged to the healthy condition interval value (results of the sugarcane assessment). In addition, the comparison of the other spectral index values (NDBaI, NDWI, and NDDI) was appropriate, as shown in Fig. 7.
- In region D, in April 2019, the NDVI value decreased, signaling an unhealthy condition (Table II). Compared with other spectral data, the value of the NDWI spectral index increased; this showed that region D in April was excessively moist. However, this condition normalized in the following month, as shown in Fig. 7.

TABLE II. NDVI VALUE IN THE MONITORING FOCUS PHASE. VALUES IN RED INDICATE AN UNHEALTHY CONDITION

Year/Month		Region			
		A	B	C	D
2019	Jan	-	-	-	-
	Feb	0.749	0.730	0.709	0.728
	Mar	0.761	0.764	0.721	0.746
	Apr	0.733	0.727	0.698	0.451
	May	0.729	0.739	0.696	0.693
	Jun	0.713	0.716	0.672	0.664
2020	Jan	0.642	0.715	0.629	0.646
	Feb	0.675	0.719	0.711	0.696
	Mar	0.734	0.768	0.693	0.756
	Apr	0.636	0.706	0.702	0.679
	May	0.685	0.728	0.458	0.698
	Jun	0.663	0.659	0.368	0.667

- In region C, in May–June 2020, the NDVI value decreased until near the harvest period (Table II), indicating the sugarcane plants were in an unhealthy condition. Together with the other spectral index values (the NDWI values decreased, while the NDDI values increased), this showed that the unhealthy condition of region C sugarcane in May–June was due to lack of water, as shown in Fig. 7.

7.932S, 112.187E | Elevation: 376 m | Climate Class: Am | Years: 2015-2019

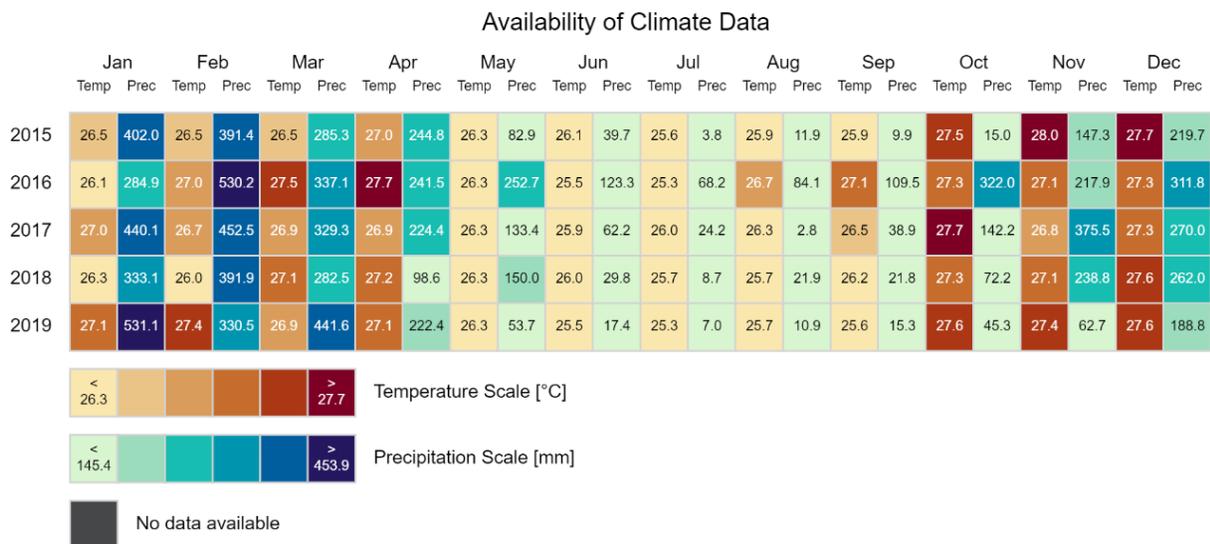


Fig. 6. Rainfall climate data in the study area from 2015 to 2019 [28]

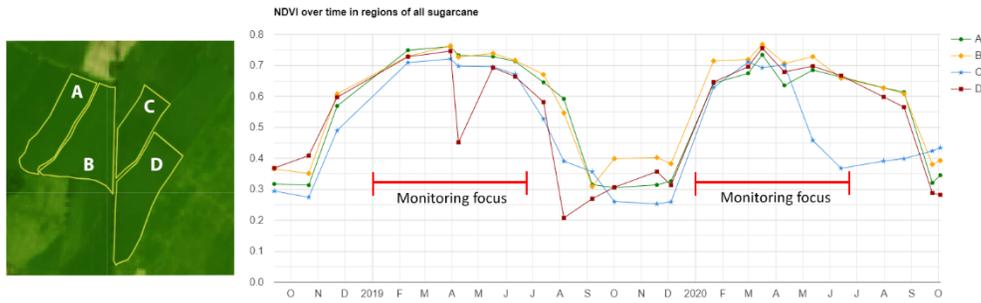


Fig. 7. The monitoring focus from January to June. The graph results of the sugarcane conditions are based on the NDVI value

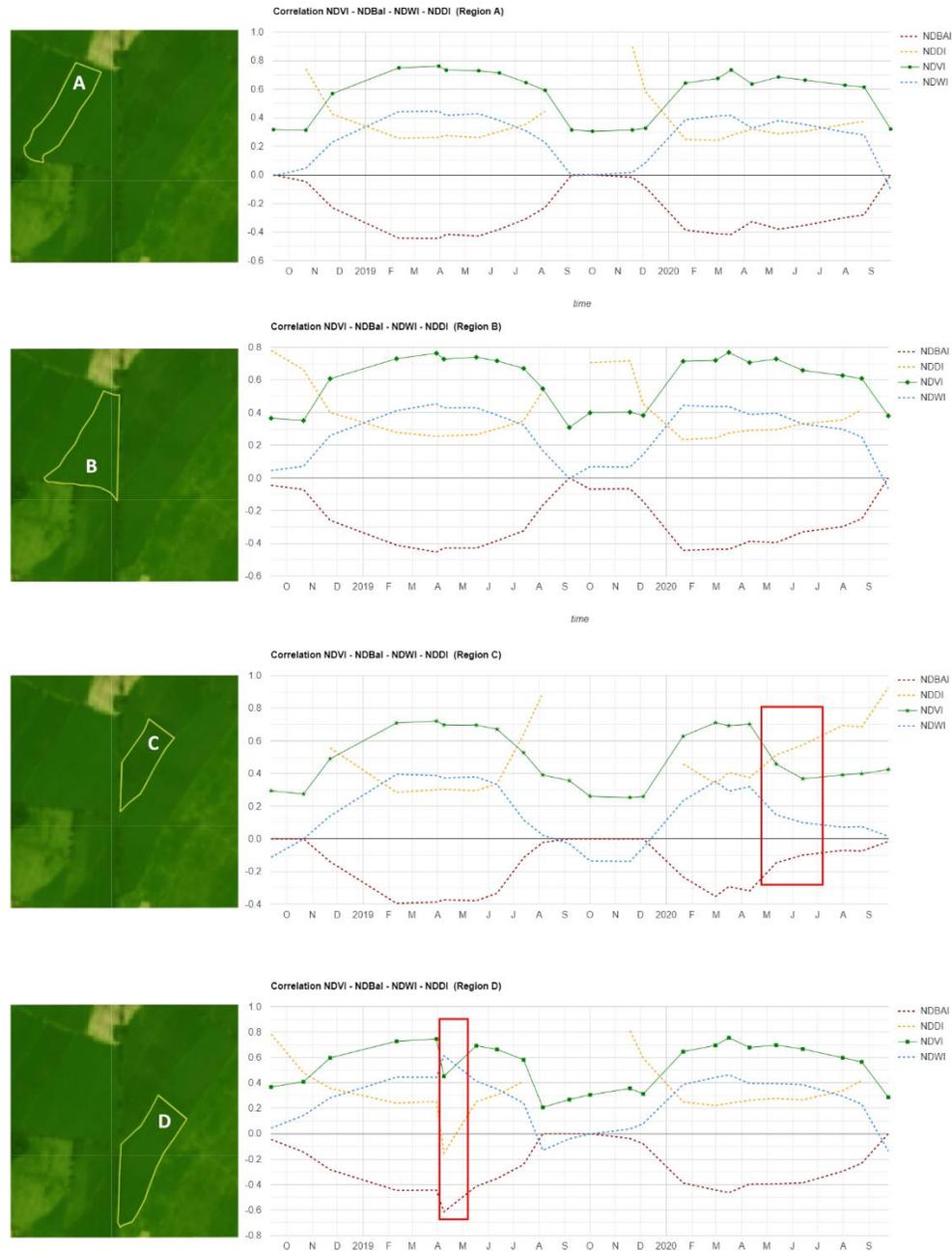


Fig. 8. Comparison of condition monitoring with several spectral index values of NDVI, NDBaI, NDWI, and NDDI from all regions. The red box is a marker showing when the planting phase was in an unhealthy condition

#### D. Discussions

The early warning was beneficial to find the factors that affected the growth of sugarcane plants. An early warning was developed based on a crucial phenological time. Phenology helps to obtain necessary information on plants, such as detecting, classifying, and monitoring plants [17]–[19], [28]–[30]. An early warning based on crucial phenology can anticipate problems with sugarcane crops and suggest practices to increase sugarcane productivity and production. In addition, early warning at crucial phenological times also provides knowledge concerning general sugarcane conditions. Although the primary index is only NDVI, it can describe the condition of sugarcane plantations when combined with other additional indices such as NDBaI, NDWI, and NDDI, unlike previous studies that only applied NDVI [5], [6], or estimates of sugarcane health conditions [8] that ignored sugarcane phenological time.

The early warning results based on phenological time show that the four regions (A, B, C, and D) had different challenges. Both regions A and B were in a healthy condition, while region D in April 2019 and region C in May–June 2020 were unhealthy. The four regions are in adjacent planting areas, but the four regions have different terrain. Regions A and B are on flat terrain, while regions C and D have sloping field characteristics.

Based on verification from other indices (NDBaI, NDWI, and NDDI), sugarcane in region D in April 2019 experienced a decrease in NDVI and NDDI value while NDWI rose, indicating that region D was waterlogged. To ensure the correct assessment of the sugarcane condition, we added a rainfall indicator in the study area [31], as shown in Fig. 8. In region D for April 2019, rainfall was still common, and in March 2019, the rainfall was relatively high. In region D, April 2019 had the potential to be waterlogged. However, this returned to normal by the following month.

In region C, May–June 2020, the sugarcane was in a drought condition. Rainfall data showed low rainfall duration and intensity, as shown in Fig. 8. Sloping terrain makes it essential for region C to apply early warning as a solution to anticipate adverse conditions. When discussed with farmers, sugarcane-growing regions need to improve water irrigation management, especially for regions characterized by sloping land. Irrigation improvement is in process with a trial of drip irrigation in August 2020.

This early warning showed factors that affected the condition of sugarcane crops. The practice of this approach to research helps early warnings become more considerate to ensure dynamic plant conditions. So solving the problem of early warning of plant growth becomes more certain against uncertain conditions. However, a drawback, namely the limited once-per-month data, results in the daily details of the sugarcane condition remaining unknown.

#### IV. CONCLUSION

Factors affecting sugarcane growth were identified using an early warning, which can help anticipate adverse conditions. In this study, it has been shown that monitoring the condition of sugar cane as an early warning, based on the phenology from

January to June, describes the condition of the sugar cane using various indicators. The sugarcane conditions in each region were different. Regions A and B were in a healthy condition throughout the monitoring phase. In contrast, regions C and D were in unhealthy conditions for part of the time. Region D had excess water in April 2019, and Region C had a drought in May–June 2020. Conditions unhealthy are based on vegetation values below 0.489. As for health conditions, it is at the opposite vegetation value.

Suggestions for future research, early warning can be solved by sharp satellite, which has a sharper resolution with a range of data available daily. In addition, early warning indicators can be combined with data related to climate change.

#### ACKNOWLEDGMENT

We thank the Institut Teknologi Telkom Purwokerto, and Department of Computer Science, IPB University for supporting this research. We also thank the anonymous reviewers for their constructive comments and advice.

#### REFERENCES

- [1] Kementerian Pertanian, “Statistik perkebunan Indonesia 2018-2020,” Buku Statistik Perkebunan Indonesia, pp. 1–82, 2020.
- [2] E. Respati, Outlook Komoditas Perkebunan Tebu. Indonesia: Pusat Data dan Sistem Informasi Pertanian Sekretaris Jenderal-Kementerian Pertanian, 2019.
- [3] G. McDonald and S. Meylinah, “Indonesia Sugar Annual Report 2019,” pp. 1–10, 2019.
- [4] T. M. Susantoro, K. Wikantika, A. B. Harto, and D. Suwardi, “Monitoring sugarcane growth phases based on satellite image analysis (A case study in indramayu and its surrounding, West Java, Indonesia),” HAYATI Journal of Biosciences, vol. 26, no. 3, pp. 117–128, 2019, doi: 10.4308/hjb.26.3.117.
- [5] T. M. Susantoro, K. Wikantika, A. Saepuloh, and A. H. Harsolumakso, “Selection of vegetation indices for mapping the sugarcane condition around the oil and gas field of North West Java Basin, Indonesia,” IOP Conference Series: Earth and Environmental Science, vol. 149, no. 1, pp. 0–10, 2018, doi: 10.1088/1755-1315/149/1/012001.
- [6] M. Rahman, A. Islam, and M. Rahman, “NDVI derived sugarcane area identification and crop condition assessment,” Plan Plus, 2004.
- [7] H. Lin, J. Chen, Z. Pei, S. Zhang, and X. Hu, “Monitoring sugarcane growth using ENVISAT ASAR data,” IEEE Transactions on Geoscience and Remote Sensing, vol. 47, no. 8, pp. 2572–2580, 2009, doi: 10.1109/TGRS.2009.2015769.
- [8] J. Morel, P. Todoroff, A. Bégué, A. Bury, J. F. Martiné, and M. Petit, “Toward a satellite-based system of sugarcane yield estimation and forecasting in smallholder farming conditions: A case study on reunion island,” Remote Sensing, vol. 6, no. 7, pp. 6620–6635, 2014, doi: 10.3390/rs6076620.
- [9] I. P. Lisboa et al., “Prediction of sugarcane yield based on NDVI and concentration of leaf-tissue nutrients in fields managed with straw removal,” Agronomy, vol. 8, no. 9, 2018, doi: 10.3390/agronomy8090196.
- [10] L. K. Petersen, “Real-time prediction of crop yields from MODIS relative vegetation health: A continent-wide analysis of Africa,” Remote Sensing, vol. 10, no. 11, pp. 1–31, 2018, doi: 10.3390/rs10111726.
- [11] N. Gorelick, M. Hancher, M. Dixon, S. Ilyushchenko, D. Thau, and R. Moore, “Google Earth Engine: Planetary-scale geospatial analysis for everyone,” Remote Sensing of Environment, vol. 202, no. 2016, pp. 18–27, 2017, doi: 10.1016/j.rse.2017.06.031.
- [12] M. Amani et al., “Google Earth Engine Cloud Computing Platform for Remote Sensing Big Data Applications: A Comprehensive Review,” IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 13, pp. 5326–5350, 2020, doi: 10.1109/JSTARS.2020.3021052.

- [13] M. Wang et al., "Land Use Policy Mapping sugarcane in complex landscapes by integrating multi-temporal Sentinel-2 images and machine learning algorithms," *Land Use Policy*, vol. 88, no. April, p. 104190, 2019, doi: 10.1016/j.landusepol.2019.104190.
- [14] L. Yang, L. Wang, and G. A. Abubakar, "High-Resolution Rice Mapping Based on SNIC Segmentation and Multi-Source Remote Sensing Images," *Remote Sensing*, 2021.
- [15] C. Luo et al., "Using time series sentinel-1 images for object-oriented crop classification in google earth engine," *Remote Sensing*, vol. 13, no. 4, pp. 1–19, 2021, doi: 10.3390/rs13040561.
- [16] C. Padwick, M. Deskevich, F. Pacifici, and S. Smallwood, "WorldView-2 pan-sharpening," *American Society for Photogrammetry and Remote Sensing Annual Conference 2010: Opportunities for Emerging Geospatial Technologies*, vol. 2, pp. 740–753, 2010.
- [17] M. Boschetti, D. Stroppiana, P. A. Brivio, and S. Bocchi, "Multi-year monitoring of rice crop phenology through time series analysis of MODIS images," *International Journal of Remote Sensing*, vol. 30, no. 18, pp. 4643–4662, 2009, doi: 10.1080/01431160802632249.
- [18] J. M. Pena-Barragan, M. K. Ngugi, R. E. Plant, and J. Six, "Object-based crop identification using multiple vegetation indices, textural features and crop phenology," *Remote Sensing of Environment*, vol. 115, no. 6, pp. 1301–1316, 2011, doi: 10.1016/j.rse.2011.01.009.
- [19] T. Sakamoto, M. Yokozawa, H. Toritani, M. Shibayama, N. Ishitsuka, and H. Ohno, "A crop phenology detection method using time-series MODIS data," *Remote Sensing of Environment*, vol. 96, no. 3–4, pp. 366–374, 2005, doi: 10.1016/j.rse.2005.03.008.
- [20] J. H. Cock, "Sugarcane growth and development," *International Sugar Journal*, vol. 105, no. 1259, pp. 540–552, 2003.
- [21] J. A. Gamon et al., "NDVI canopy structure photosynthesis," vol. 5, no. 1, pp. 28–41, 1995.
- [22] J. W. Rouse, R. H. Haas, J. A. Schell, and D. W. Deering, "Monitoring the vernal advancement and retrogradation (green wave effect) of natural vegetation," *Progress Report RSC 1978-1*, p. 112, 1973.
- [23] H. Zhao and X. Chen, "Use of normalized difference bareness index in quickly mapping bare areas from TM/ETM+," *International Geoscience and Remote Sensing Symposium (IGARSS)*, vol. 3, no. December 2017, pp. 1666–1668, 2005, doi: 10.1109/igarss.2005.1526319.
- [24] B.-C. Gao, "NDWI A Normalized Difference Water Index for Remote Sensing of Vegetation Liquid Water From Space," *Online Learning*, vol. 23, no. 3, pp. 257–266, 1996, doi: 10.24059/olj.v23i3.1546.
- [25] Y. Gu, J. F. Brown, J. P. Verdin, and B. Wardlow, "A five-year analysis of MODIS NDVI and NDWI for grassland drought assessment over the central Great Plains of the United States," *Geophysical Research Letters*, vol. 34, no. 6, pp. 1–6, 2007, doi: 10.1029/2006GL029127.
- [26] D. S. Shumway, R.H., & Stoffer, *Time Series: A Data Analysis Approach Using R: A Data Analysis Approach Using R*. 2019.
- [27] NASA Science, "Reflected Near-Infrared Waves | Science Mission Directorate," Nasa. 2018.
- [28] V. Simonneaux, B. Duchemin, D. Helson, S. Er-Raki, A. Olioso, and A. G. Chehbouni, "The use of high-resolution image time series for crop classification and evapotranspiration estimate over an irrigated area in central Morocco," *International Journal of Remote Sensing*, vol. 29, no. 1, pp. 95–116, 2008, doi: 10.1080/01431160701250390.
- [29] B. D. Wardlow and S. L. Egbert, "Large-area crop mapping using time-series MODIS 250 m NDVI data: An assessment for the U.S. Central Great Plains," *Remote Sensing of Environment*, vol. 112, no. 3, pp. 1096–1116, 2008, doi: 10.1016/j.rse.2007.07.019.
- [30] P. Serra and X. Pons, "Monitoring farmers' decisions on Mediterranean irrigated crops using satellite image time series," *International Journal of Remote Sensing*, vol. 29, no. 8, pp. 2293–2316, 2008, doi: 10.1080/01431160701408444.
- [31] L. Zepner, P. Karrasch, F. Wiemann, and L. Bernard, "ClimateCharts.net—an interactive climate analysis web platform," *International Journal of Digital Earth*, vol. 14, no. 3, pp. 338–356, 2021, doi: 10.1080/17538947.2020.1829112.

# Fault Tolerance Smart Egg Incubation System with Computer Vision

Emilijan Petkov, Teodor Kalushkov, Donika Valcheva, Georgi Shipkovenski

Department of Mathematics and Informatics, University of Veliko Tarnovo, Veliko Tarnovo, Bulgaria

**Abstract**—Reliability of incubators is one of their most important specifications. Development of wireless, cloud and computer vision based technologies gives new possibilities for work process control and increasing fault tolerance. Regardless of whether the hatching is in the field of mass production or in the breeding of rare species of birds, detecting a critical situation and sending timely notifications can prevent serious losses. Experience shows that network isolated solutions are not reliable enough and good management requires complex algorithms that are beyond the capabilities of a local, single controller. Even with the duplication of some sensors and actuators, incubators without external connection are high risk due to the fact that their controller is a central point in the architecture and can fail, leaving the farmer with no alert about the accident. The report presents a solution that uses periodic checks from cloud structures on the condition and operability of the incubator. In parallel, a video surveillance system analyzes the internal environment and the condition of hatching chicks. When potential and real risks occur, the system sends notifications to the responsible persons even to his or her wrist. Additionally, the proposed smart egg incubation methodology has been found to reduce the amount of time required for farmers to oversee the incubation process by up to 50%, allowing them to focus on other important tasks while still ensuring optimal hatching conditions for their eggs. Overall, the proposed methodology offers a significant improvement in egg incubation efficiency and reliability, with potential applications in both commercial and personal settings.

**Keywords**—Hatching; incubation; computer vision; cloud architecture; sending alerts; smart farming; internet of things

## I. INTRODUCTION

### A. Defining the Research Problem

Smart farming, like industrial production, follows the steady trend of replacing manual human labor with automated systems. Innovative solutions manage not only purely mechanical activities but also digital ones implementing artificial intelligence and computer vision in many systems to generate feedback for control and corrections of processes during their operation. In this way, the possibility of human error is reduced to the lowest possible values and the dependence on the labor market is minimized. The ultimate goal that is achieved through these trends is a significant increase in efficiency and respectively productivity. First smart farming solutions that have been developed include systems with controllers, sensors and actuators, but without video cameras.

The prerequisites for the development of computer vision based smart farming include:

- Offering cameras with high matrix resolutions and the ability to capture many frames per second on a lower price.
- Development of new network standards for high-throughput wireless communications.
- More accessible cloud technologies that provide the ability to process more complex algorithms in image analysis.
- Ability to customize processing load distribution between on-premises and cloud structures.
- The trends for a non-expansive increase in a farm production.

It is a hard task to list all the examples in this area, but some of them may be referred to as a representative sample. In crop production, a large group of computer vision based applications aims to recognize the fruits of plants in order to pick them [1,2]. Some others analyze pictures of plant leaves to find diseases and determine their condition [3,4,5]. The outcomes of such systems can be used for appropriate subsequent treatment (spraying) of the plants [6].

There are no less in a number of examples of computer vision systems in animal husbandry. Examples for animal farming include counting of silkworm eggs [7]. Sometimes in animal farms, the processes of image acquisition and processing can be significantly complicated due to the more intensive motor activity of animals compared to that of plants. Such case with pigs is described in [8]. It describes the algorithms for separating individual animals from a group in which they are close to each other. Of course there are again examples of recognizing health problems of animals from the processed images [9,10]. Feeding process control is also a possible target of the computer vision system in a smart farm [10]. This paper explains how to recognize pigs with low weight. Poultry farms are no exception to the implementation of intelligent machine vision solutions. In this scenario infertile egg detection is a key feature [11].

Some interesting applications corresponding with smart farming also exploit computer vision methods. Baoming Shan uses image analyses of eggs in vaccine production from embryos [12]. Fertility detection is again targeted in the results, although the field of use in this case is pharmacy.

A large number of papers [13,14,15,16] oriented to a human incubators describe machine-vision mechanisms for monitoring the state of babies. Different type of cameras can monitor movements, temperature and persistence of a newborn. From the recorded movements, it is possible to draw conclusions about certain behavior according to a health condition and thus detect critical situations. The described systems are traditionally used in combination with standard sensors and controllers that upgrade the information from the images.

All listed examples above lead to the conclusion that computer vision can be very helpful into the hatching process and may increase the reliability, by protecting the eggs and new chicks from injury and exposure to adverse conditions in the incubator's internal environment.

### B. Importance of the Problem

Today the share of small farms that breed several species of animals is decreasing compared to that of large and strictly specialized in certain breeding. Therefore, methods and devices for hatching a large number of chickens are sought in poultry farming. In case of hatchery failure, this approach makes the potential losses significantly greater.

Apart from hatching for mass production, with the right settings, the small, modern, intelligent hatcheries can also be used for breeding exotic species of birds and reptiles. This scenario is also associated with a high risk of failure because eggs are usually very rare.

### C. Purpose and Tasks of the Research

In order to reduce the possibility of harmful consequences in the situations described above, it is necessary to increase the reliability of the hatcheries using modern methods. This main purpose can be divided into the following tasks:

- Selection of building modules.
- Implementing cameras in the hatcheries so that operators can monitor the interior.
- Creation and implementation of a computer vision system in the hatchery.
- Preserving the functionalities for reliable notifications of the operator in a problematic situation.
- Analysis of possibilities and selection of a model for building the connections between the devices, the cloud environment and the operators.
- Implementation of a prototype corresponding to the selected model.

### D. Hypothesis

Cloud technologies can be a key tool to increase the reliability. The integration of a computer vision system can be realized locally through a controller and a smart camera, but such solutions are too expensive. On the other hand, using a microcomputer and camera is an alternative, but using an operating system can generate new potential points of failure (opportunities for viruses, attacks, sudden reboot problems, etc.). Cloud structures provide an optimal solution because

they do not require computing potential from the controllers, but possess the necessary for the implementation of the functions desired by the user.

### E. Expected Benefits

The results of research, described in this paper try to offer a model for building a modern incubator with maximal fault tolerance, level of automation, flexible control and visual feedback for the users. Suggested devices should also allow comfortable parameters observation (charts) and optimization of the processes. Following this line of thought can be defined a new system of incubators as Fault Tolerance Smart Egg Incubation System with Computer Vision (FTSEISCV). The reader should note that this is a development of another study on smart incubators. Prior publication on development of Smart Egg Incubator (SEI) can be seen at [17].

The main contributions of the proposed smart egg incubation system include advanced features such as microcontroller-based control and monitoring, camera-based hatching detection, and precise regulation of temperature, humidity, and airflow parameters, which collectively provide farmers with an efficient, reliable, and user-friendly solution for egg incubation. Additionally, the system's ability to reduce the amount of time required for farmers to oversee the incubation process by up to 50% can have a significant impact on productivity and profitability, while also improving hatching success rates and reducing the risk of losses due to incubation-related issues.

### F. Organization of the Paper

The paper contains an overview of historical development of incubators in the second part and analysis of used models up to the moment in the third. Next fourth part represents the proposed, improved model, selected hardware components for the experimental prototype, used software and cloud structures. User interface and control management are also explained in this part. The paper finishes with conclusions and references.

## II. OVERVIEW OF EXISTING SOLUTIONS ON THE TOPIC

The first incubators from 20th century were made from a wooden box with a pot-like container attached to the side (Fig. 1) [18].



Fig. 1. An old incubator from 1913.

They included a thermometer and required almost constant monitoring by farmers.

Electronic analog-element managed systems increased the automation. They included humidity and temperature sensors

and actuators like heaters, fans, evaporators and mechanisms for eggs rotation.

After IC chips implementation, some advanced and more complicated functions were able to be performed. Specific to this period is the transition from analog to digital signal processing.

Until a few years ago the only accessible option to increase reliability was realized by duplicating individual elements of the incubator. Thus, when one of the elements fails, the backup one takes over its functions. However, these solutions often do not work if the controller module fails and therefore require regular checks from operator to prevent disaster.

First notification developments were realized through mobile network gateways. At this moment were implemented alerts for the operators, when incubation parameters exceed the preset limits. Unfortunately, if the cellular network is not available or if the user is temporarily unavailable (due to personal reasons sometimes), the notifications are not repeated and thus the desired level of trust is not guaranteed.

Programmable controllers allow full automatic control during the whole incubation. In addition to temperature and humidity, sensors in modern incubators also measure the percentage content of oxygen, carbon dioxide and ammonia compounds in the air.

The limitations of existing egg incubation systems, such as poor temperature and humidity control, inconsistent egg rotation, and limited hatching success rates, motivated the development of the proposed smart egg incubation system with advanced features like microcontroller-based control and monitoring, camera-based hatching detection, and precise regulation of temperature, humidity, and airflow parameters.

The next step in the upgrade of the modern incubators is realized with the expansion of the World Wide Web and their digital transformation in the industry [19] into Internet of Things devices. Initial attempts in this direction were aimed to monitor the processes and environment and control them over the network, only by the operators. At this point, in addition to cellular networks, notifications could now be sent and received in parallel or alternatively via the Internet. This kind of device operation can be defined as a third layer of the mentioned model above.

### III. OUTCOME OF THE RESEARCH ON THE EXISTING SOLUTIONS AND A PARTICULAR CONCEPT

As an outcome of research on the existing solutions, can be made a conclusion that the first incubators were built only on one layer, namely a device level. Then the users did not have own terminal devices for monitoring and control, and also cloud structures did not exist in this period. On objective reasons, it was necessary for the operators to be constantly near to the hatcheries. Later solutions were constructed over two-layer model. It contains client and device layers, because users already own a primitive mobile device like pagers, 1G or 2G phones. Connections between levels in this case were straight and were built primarily over mobile networks.

Thinking about the possibility of incubators being connected to the Internet and controlled by client devices, such as computers and smart phones, cloud technologies were chosen as the most intelligent solution to achieve the goal set in the research. The Cloud comes in as an intermediate layer (between Device and Control) to store and process data, as well as provides services such as video streaming, notifications, routines deployment, and more. Therefore, the chosen model for creation the FTSEISCV, is an adaptation of the three-layer model known from the IoT architecture, namely Application–Network–Perception model [20,21].

In the present research an adaption to Device–Cloud–Control was realized. Device stands for the new smart incubator, Cloud – for the cloud solutions and Control – for the farmer’s computers and mobile devices.

The adapted three-layer model shown in Fig. 2 incorporates the additional cloud layer that enables the realization of the IoT concept and a wide variety of additional services as a part of next step in global network development [22]. The cloud layer mediates between user devices and the managing controller part. The connections between the layers are now realized with priority over the Internet. The three-layer model is of increased complexity, but provides functional flexibility in many ways, including improved reliability. Therefore, the three-layer model is the choice on which is based this study.



Fig. 2. Three-layer model.

Following modern trends, there is no need to try implementation of new layers, but improvement of the current ones and optionally dividing any of them into different parts. On device level except sensors, controllers and actuators, additional imaging devices can be introduced. On a cloud level a solution can be to migrate some of the functionalities in order to prevent overload in a controller part (device level) and to provide an external point of operability testing. Cloud layer also takes care to guarantee secure notification to the user.

Here are a few examples of state-of-the-art egg incubators that you can find online:

- GQF Manufacturing "Sportsman 1502" Cabinet Incubator: This large-capacity incubator can hold up to 1368 quail eggs or 270 chicken eggs. It includes a digital control system for temperature and humidity, and automatic egg turning. It also features a built-in fan for improved airflow [23].

- R-Com "20 MAX" Egg Incubator: This incubator features a microprocessor-based control system with automatic temperature and humidity adjustment. It also includes an automatic egg turning system and a built-in fan for improved airflow [24].
- Farm Innovators "Model 4250" Digital Circulated Air Incubator: This compact and affordable incubator uses a digital thermostat and fan to provide precise temperature and humidity control. It includes an automatic egg turner and a clear plastic dome for easy observation [25].

By exploring these and other state-of-the-art egg incubators available on the market, the reader can gain a better understanding of the features and capabilities that are available in today's incubation systems, and compare them with the features and capabilities of the presented here smart egg incubation system.

#### IV. DEVELOPMENT OF FTSEISCV

The development of the whole system includes creation of its conceptual model, device architecture, cloud functionalities and control applications. All devices, cloud components and applications are designed to work as one unified system. The basic part of every system is its model. So, first step is to emphasize the model of the new system.

##### A. Model

Creation of conceptual model is based on the scientific analysis, conducted and described in the previous point (Fig. 3).

It presents a number of SEI in the Device layer that connects to the cloud solutions via the Internet. On the other side are the control devices in the Control layer that communicate with the cloud solutions. As a result, the farmer has all the data and functions necessary to control the SEI.

Prototype planning details include a segmentation of the components along the layers of the adopted model and an organization of the connections between them in an optimal way. After extensive analysis and partial approvals, the model presented in details on Fig. 4 has been created.

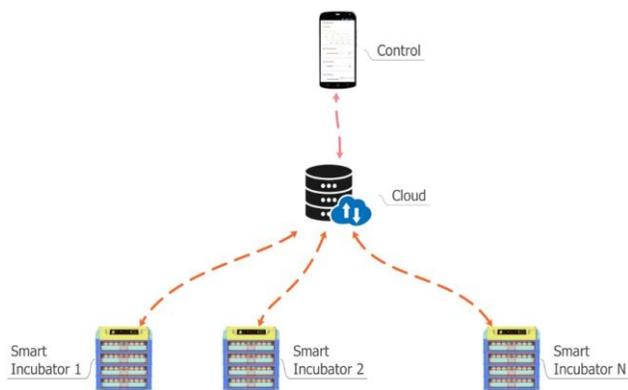


Fig. 3. Conceptual model of FTSEISCV.

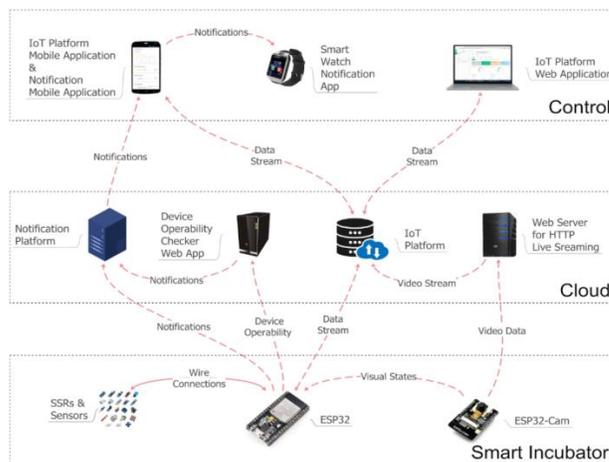


Fig. 4. Detailed model of FTSEISCV.

##### B. SEI Architecture and Components

Selection of the hardware for the FTSEISCV in the device layer is based on components that are available relatively cheap and at the same time does not compromise on their reliability and performance. The popular development environment and the support of many already created libraries for the components are other main criteria.

In practice, one of the innovations of the developed system is the provision of fault tolerance, which does not rely only on failsafe elements. On the second level, it is guaranteed by the cloud structures and the well-chosen connections between the components. Additional cloud applications ensure this functionality.

Fig. 5 represents all selected components for the incubator.

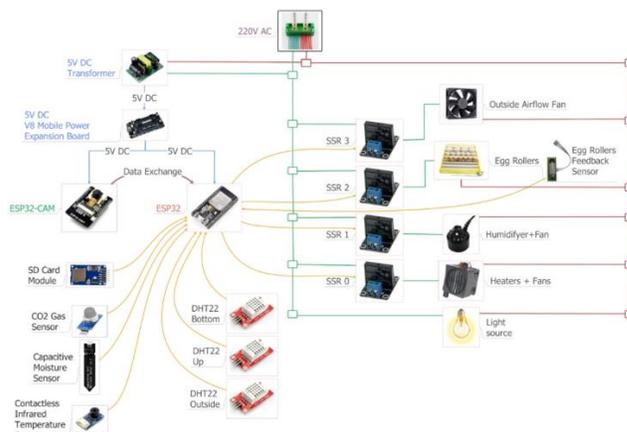


Fig. 5. SEI architecture.

There are three DHT22 temperature and humidity sensors. They fully cover the requirements for accuracy, working range, consumption and supplying digital values to the controller. Two are located inside to monitor the homogeneity of the hatching environment. The third one is placed outside to increase the attention of the operator in the case of too great differences in the external environment compared to the internal one, because in such situations they imply increased risks and overloads.

Capacitive moisture sensor located outside works as a level meter gauge in the outer water container of humidifier, which is used to increase the humidity if necessary. Capacitive humidity sensors are more suitable than resistive ones, because when measuring the resistance between the individual electrodes electrolysis occurs and they corrode and fail.

Contactless infrared temperature sensor is placed inside and is directed at the shell of one of the eggs. It acts as system feedback, supplementing the readings of the DHT22 sensors and increases reliability.

The CO<sub>2</sub> gas sensor is most actively used in the hatching phase and aims to notify the farmer of a potential suffocation danger to the chicks. When they have already hatched, they become active, breathe rapidly, releasing carbon dioxide accordingly.

Another sensor is positioned on the egg rollers and provides feedback for rotation process. It is an encoder that switches its logic states between of zero and one with a certain frequency determined by the speed of rotation. To work reliably, it must be well sealed and waterproof.

The SD Card Module shown on the schematic is optional and can be used to store the data read from the sensors before being sent to the cloud database. When this module is not used, these functions can be performed by pre-allocated partitions of controller memory.

As mentioned above, the controller occupies a central place in the hardware architecture. That is why its choice is one of the most important to achieve the planned global goals. By using Texas Instruments or Nordic Semiconductors devices for example, can be achieved really good final solutions, but the value of the incubator in these cases will increase significantly. On the other hand, NodeMCU (ESP8266) is also a possible solution, but the potential for future development and its lower reliability when performing more functions; it is not the best choice. Compared to it, ESP32 offers not one but two computing cores, double the clock speed, more input-output pins, more SRAM memory and an additional Bluetooth channel for communication in parallel with WiFi. ESP32 is also suitable because of its ability to use the ArduinoIDE programming environment, which is free and has a wide variety of already developed and available libraries.

To implement the computer vision system, it is necessary to select a suitable camera. Variants with a USB connection are not suitable, because the ESP32 controller will have to transmit simultaneously the video stream and data from the sensors via WiFi. This can create conflicts and competition for the communication channel and thus reduce reliability. The purchase of smart cameras with their own computer vision system and/or Wi-Fi connection to the cloud structures is an inappropriate optional solution because it can again lead to an excessive increase in price and for one reason or another, not covers the requirements for full compatibility between all system components. In accordance with the stated requirements, the best choice remains ESP32 Cam. Through its Wi-Fi channel that is parallel and independent compared to that of the ESP32 controller, the camera can stream video to the network. If direct communication with the controller is

required for data transmission, one of the additional GPIOs or the Bluetooth connection can be used. An additional functionality that may not be used is the SD card slot with which the camera is equipped. The ESP32 Cam is available on the market with variable focal length lenses that are selected according to the specific internal hatchery design.

AC/DC power supply 2A 5V is used for the controller, camera and sensors. Its rated power significantly exceeds the total power consumed. An additional V8 mobile power expansion board with its own battery is selected to protect the controller and camera from unwanted reboots. It operates as UPS in the event of a central power failure. Thanks to it, in such a critical situation, the system can continue to monitor the readings from the sensors and send data if there is an available access point, even though the actuator parts are not active.

The actuators on the Fig. 5 are light source, outside airflow fan, egg rollers, humidifier with fan and heaters with fan. They all use 220V mains power supply. Heaters maintain the necessary internal temperature, and their fan aims to spread the heat evenly inside. In conditions of excessive heating or too high humidity, the internal environment in the hatchery is normalized by means of the external airflow fan.

In contrast to heaters, practical tests have shown that humidifier should be placed outside the incubator. Its fan conducts the moisture evaporated from the water through specially designed pipes to the interior.

Egg rollers are placed on a common frame and are rotated simultaneously by a stepper motor attached to a worm gear mechanism. The position of its shafts can be adjusted according to the size of the eggs being incubated.

Every computer vision system needs appropriate lighting source. Therefore, the camera in SEI needs lighting with a stable intensity to work correctly. In the experimental prototype, this role is performed by a constantly glowing 7W LED lamp.

The connection between the controller and the actuators is performed by 4 electronic solid state relays. They are significantly more reliable than electro-mechanical ones because they do not have the possibility of the contacts self-welding during operation.

### C. Functional Capabilities

Blynk cloud solution has been chosen for the practical implementation of the system. Blynk allows registration of multiple devices. In this particular case, the SEI software that works with the main ESP32 controller and the auxiliary ESP32 Cam was developed. Blynk supports the device with proper version of the software automatically. The software communicates with Blynk to transmit the values from the sensors, the states in which the incubators are at a particular moment, the video stream and others.

The system is designed to be controlled by the farmer's control devices (computer and mobile phone). For this purpose graphical user interface in Blynk application has been developed for mobile devices and computers (Fig. 6). It allows the user to start, stop and pause the incubators, to set the values of temperature, humidity, critical temperatures, critical

humidity, etc. that the incubators have to control, to watch the video stream from inside the incubator and many others. The GUI allows the user to send commands to the SEIs through a terminal window in order to control them or to receive specific information for their states (Fig. 7).

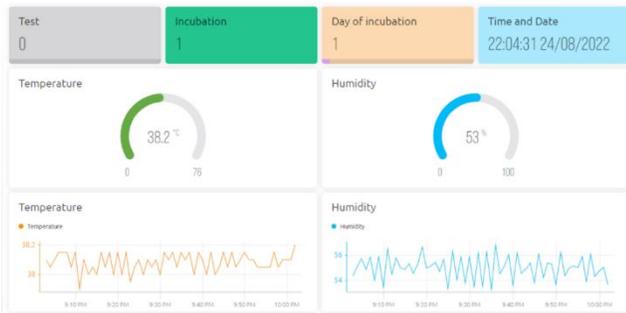


Fig. 6. Graphical user interface in Blynk.

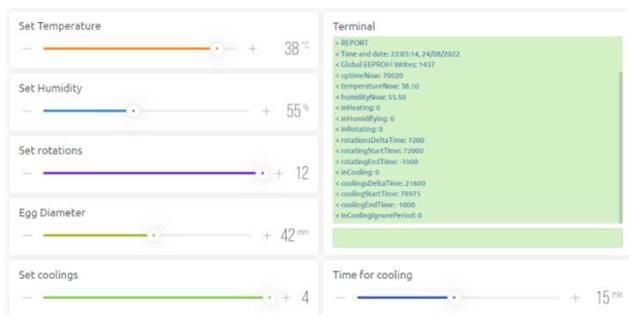


Fig. 7. Terminal window in Blynk.

Telegram has been chosen as a notification platform. When critical notifications appear, they are sent from the SEIs through Telegram to the mobile phone and eventually to the smart watch of the farmer. At that moment the user can react immediately to prevent damage of the production.

A special application has been created to check the operability of the devices at any time. If it detects a problem with a device it sends a proper notification through Telegram to the farmer.

An algorithm for vision inspection of the eggs has been developed and implemented to work in the ESP Cam modules. Its purpose is to detect the moment of hatching a chicken from an egg. If the module detects such an event, it sends a proper notification to the farmer. Then the farmer can observe the incubator on the video stream or at place (Fig. 8).

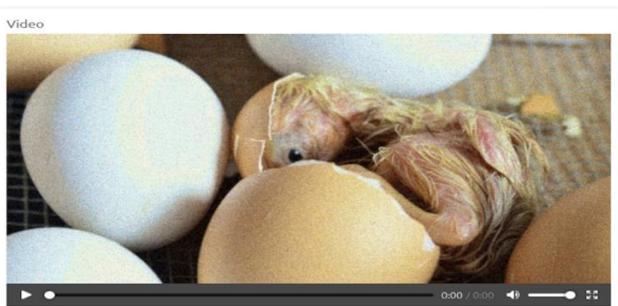


Fig. 8. Video stream interface through the IoT cloud platform.

## V. CONCLUSIONS AND FUTURE WORK

A real prototype has been assembled for approbation of the proposed FTSEISCV model. Partial and complete tests show stable results, proving a high degree of reliability, functional conveniences and a high degree of automation that guarantees a minimal commitment of users to the incubation process. By multiplying through a clustered approach, the system can be used on an industrial scale without compromising the intended performance.

Further researches will be focused on finding an optimal position of multiple cameras for the target area, management of the parallel video streams, improvement of computer vision algorithm and implementing of proactive features in the incubation control.

## REFERENCES

- [1] A. B. Titus, T. Narayanan and G. P. Das, "Vision system for coconut farm cable robot," 2017 IEEE International Conference on Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials (ICSTM), Chennai, India, 2017, pp. 443-450, doi: 10.1109/ICSTM.2017.8089201.
- [2] H. Xiaomei, L. Shunke and C. Jianfei, "Design of Agaricus Bisporus Smart Farm," 2019 IEEE 4th International Conference on Image, Vision and Computing (ICIVC), Xiamen, China, 2019, pp. 351-355, doi: 10.1109/ICIVC47709.2019.8981061.
- [3] H. Phan, A. Ahmad and D. Saraswat, "Identification of Foliar Disease Regions on Corn Leaves Using SLIC Segmentation and Deep Learning Under Uniform Background and Field Conditions," in IEEE Access, vol. 10, pp. 111985-111995, 2022, doi: 10.1109/ACCESS.2022.3215497.
- [4] L. Feng et al., "Multitask Learning of Alfalfa Nutritive Value From UAV-Based Hyperspectral Images," in IEEE Geoscience and Remote Sensing Letters, vol. 19, pp. 1-5, 2022, Art no. 5506305, doi: 10.1109/LGRS.2021.3079317.
- [5] S. Kumar, G. Chowdhary, V. Udutalappally, D. Das and S. P. Mohanty, "gCrop: Internet-of-Leaf-Things (IoLT) for Monitoring of the Growth of Crops in Smart Agriculture," 2019 IEEE International Symposium on Smart Electronic Systems (iSES) (Formerly iNiS), Rourkela, India, 2019, pp. 53-56, doi: 10.1109/iSES47678.2019.00024.
- [6] B. -X. Xie, C. -H. Wang, J. -Y. Ke and C. -L. Chang, "Design and Implementation of a Machine Visionbased Spraying Technique for Smart Farming," 2020 International Automatic Control Conference (CACs), Hsinchu, Taiwan, 2020, pp. 1-5, doi: 10.1109/CACs50047.2020.9289789.
- [7] A. Pandit, J. Rangole, R. Shastri and S. Deosarkar, "Vision system for automatic counting of silkworm eggs," International Conference on Information Communication and Embedded Systems (ICICES2014), Chennai, India, 2014, pp. 1-5, doi: 10.1109/ICICES.2014.7034036.
- [8] J. Seo, J. Sa, Y. Choi, Y. Chung, D. Park and H. Kim, "A YOLO-based Separation of Touching-Pigs for Smart Pig Farm Applications," 2019 21st International Conference on Advanced Communication Technology (ICACT), PyeongChang, Korea (South), 2019, pp. 395-401, doi: 10.23919/ICACT.2019.8701968.
- [9] S. Lee, H. Ahn, J. Seo, Y. Chung, D. Park and S. Pan, "Practical Monitoring of Undergrown Pigs for IoT-Based Large-Scale Smart Farm," in IEEE Access, vol. 7, pp. 173796-173810, 2019, doi: 10.1109/ACCESS.2019.2955761.
- [10] I. -J. Huang et al., "The Prototype of a Smart Underwater Surveillance System for Shrimp Farming," 2018 IEEE International Conference on Advanced Manufacturing (ICAM), Yunlin, Taiwan, 2018, pp. 177-180, doi: 10.1109/AMCON.2018.8614976.
- [11] L. K. S. Tolentino, E. Justine G. Enrico, R. L. M. Listanco, M. Anthony M. Ramirez, T. L. U. Renon and M. Rikko B. Samson, "Development of Fertile Egg Detection and Incubation System Using Image Processing and Automatic Candling," TENCON 2018 - 2018 IEEE Region 10 Conference, Jeju, Korea (South), 2018, pp. 0701-0706, doi: 10.1109/TENCON.2018.8650320.

- [12] B. Shan, "Fertility Detection of Middle-stage Hatching Egg in Vaccine Production Using Machine Vision," 2010 Second International Workshop on Education Technology and Computer Science, Wuhan, China, 2010, pp. 95-98, doi: 10.1109/ETCS.2010.540.
- [13] Y. S. Dosso, R. Selzler, K. Greenwood, J. Harrold and J. R. Green, "RGB-D Sensor Application for Non-Contact Neonatal Monitoring," 2021 IEEE Sensors Applications Symposium (SAS), Sundsvall, Sweden, 2021, pp. 1-6, doi: 10.1109/SAS51076.2021.9530044.
- [14] Z. Peng et al., "A Comparison of Video-based Methods for Neonatal Body Motion Detection," 2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), Glasgow, Scotland, United Kingdom, 2022, pp. 3047-3050, doi: 10.1109/EMBC48229.2022.9871700.
- [15] K. Rassels and P. French, "Accurate Body Temperature Measurement of a Neonate Using Thermography Technology," 2021 Smart Systems Integration (SSI), Grenoble, France, 2021, pp. 1-5, doi: 10.1109/SSI52265.2021.9467024.
- [16] R. Weber, S. Cabon, A. Simon, F. Porée and G. Carrault, "Preterm Newborn Presence Detection in Incubator and Open Bed Using Deep Transfer Learning," in IEEE Journal of Biomedical and Health Informatics, vol. 25, no. 5, pp. 1419-1428, May 2021, doi: 10.1109/JBHI.2021.3062617.
- [17] E. Petkov, T. Kalushkov, G. Shipkovenski, R. Radoeva and D. Valcheva, "Fault Tolerance Smart Incubator With IoT Control and Alerts," 2022 International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT), Ankara, Turkey, 2022, pp. 929-933, doi: 10.1109/ISMSIT56059.2022.9932792.
- [18] <https://www.kijiji.ca/v-art-collectibles/stratford-on/buckeye-incubator-1913-excellent-instruction-manual-thermo/> last accessed 2022/12/21.
- [19] K. K. Angelov, S. M. Sadinov, and P. Kogias, "Deployment of mesh network in an indoor scenario for application in IoT communications." The International Conference on Communications Information Electronic and Energy Systems (CIEES 2022) IOP Conference Series Materials Science and Engineering 1032, vol. 1, pp. 012004 (2020).
- [20] <https://www.elprocus.com/iot-protocols-and-its-architectures/> last accessed 2023/01/05.
- [21] M. R. Abdmeziem, D. Tandjaoui, I. Romdhani, "Architecting the Internet of Things: State of the Art" In book: Robots and Sensor Clouds, Special edition in the "Studies in Systems, Decision and Control" (pp.55-75) Edition: Springer International Publisher: Springer Book Series (2016).
- [22] H. Beloev, A. Smrikarov, A. Ivanova, T. Vassilev, T. Georgiev, S. Smrikarova, G. Ivanova, V. Stoykova, E. Ibryamova, Y. Aliev, and P. Zlatarov, "A Vision of the University of the Future." In: Proceedings of the 21st International Conference on Computer Systems and Technologies' 20, pp 307-312 (2020).
- [23] <https://incubatorwarehouse.com/hova-bator-cabinet-egg-incubator-1502.html> last accessed 2023/01/11.
- [24] <https://rcom-incubator.com/products/r-com-mx-20-incubator> last accessed 2023/01/11.
- [25] <https://www.farminnovators.com/4250i-sheet2018.pdf> last accessed 2023/01/11.

# A Novel Hybrid Deep Learning Framework for Detection and Categorization of Brain Tumor from Magnetic Resonance Images

Yousef Methkal Abd Algani<sup>1</sup>, Dr. B. Nageswara Rao<sup>2</sup>, Dr. Chamandeep Kaur<sup>3</sup>, B. Ashreetha<sup>4</sup>, K.V.Daya Sagar<sup>5</sup>,  
Prof. Ts. Dr. Yousef A.Baker El-Ebiary<sup>6</sup>

Department of Mathematics, The Arab Academic College for Education in Israel-Haifa, Israel<sup>1</sup>

Associate Professor, Department of Science and Humanities, Lendi Institute of Engineering and Technology, Jonnada,  
Vizinagaram<sup>2</sup>

Lecturer, Dept of IT, Jazan University, Saudi Arabia<sup>3</sup>

Assistant Professor, Department of Electronics & Communication Engineering College-Sree Vidyanikethan Engineering College,  
Mohan Babu University, Tirupati, Andhra Pradesh, India<sup>4</sup>

Associate Professor, Electronics and Computer Science, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur,  
Andhra Pradesh, India<sup>5</sup>

Faculty of Informatics and Computing, UniSZA University, Malaysia<sup>6</sup>

**Abstract**—Cellular abnormality leads to brain tumour formation. This is one of the foremost reasons of adult death all over the world. The typical size of a brain tumour increases within 25 days due to its rapid growth. Early brain tumour diagnosis can save millions of lives. For the purpose of early brain tumour identification, an automatic method is necessary. MRI brain tumour detection improves the survival of patients. Tumour visibility is improved in MRI, which facilitates subsequent treatment. To distinguish between brain MRI images with tumour and images without tumour is suggested in this paper. Many approaches in the field of machine learning including Support Vector Machine, Artificial Neural Networks, and KNN classifier have been developed for solving these issues. But these methods are time consuming, inefficient and require complex procedures. For a Computer Assisted Diagnosis system to aid physicians and radiologists in the identification and categorization of tumours, artificial intelligence is used. Deep learning has demonstrated an encouraging efficiency in computer vision systems over the past decade. In this paper, identification and classification of brain tumour from MR images employing BGWO-CNN-LSTM method is proposed. The proposed method on a testing set with 6100 MRI images of four different kinds of brain tumours is utilized. In comparison to earlier research on the same data set, the suggested approach achieved 99.74% accuracy, 99.23% recall and 99.54% specificity which are greater than the other techniques.

**Keywords**—Brain tumour; BGWO; LSTM; CNN; MRI dataset

## I. INTRODUCTION

The brain is a massive, intricate part that governs the entire nervous system in humans and comprises an estimated hundred billion nerve cells. This vital organ was grown in the brain's cerebral cortex. As a result, any brain abnormalities could be dangerous for people's health. Brain tumours are among the most serious of these disorders [1]. For the proper operation of the human body, the majority of the cells which are produced in the body split to become new cells. Older or

damaged biological cells expire as new ones proliferate. New cells then replace the old ones. Occasionally, when the body does not require them, new cells are produced. Extra cells produced in the body build a form of tissue known as a tumour. The delicate bodily functioning is distorted by a tumour positioned in the brain area. Due to its position and tendency to spread, it is extremely dangerous and difficult to cure [2]. Tumors can be classified as either cancerous as well as non-cancerous. Both benign and malignant brain tumours are recognized as dangerous conditions. The tumour spreads throughout the skull, flattening the other development. Primary tumours and secondary tumours are the two different categories of brain tumours. The primary tumour arises in the tissues of the brain, and the secondary tumour spreads from other body regions to the skull [3]. Pituitary tumours, gliomas and meningiomas are a few kinds of primary brain tumours. Tumors that form in brain tissues apart from nerve cells and blood vessels are referred to as gliomas in context. Pituitary tumours are lumps that reside within the skull; however meningiomas safeguard and surround the central nervous system and the brain. [4]. The meninges covers the brain within the skull, is where meningioma tumours first appear. These tumours are benign because of their sluggish growth. The malignant tumours known as gliomas begin in the brain, spinal cords, and nearby glial cells, which are nerve cells [5]. Developing an efficient therapy for brain tumours depends on an early and precise diagnosis of the condition. The pathologic grade, type, and tumour stage at the point of diagnosis significantly influence the treatment mode selection [6]. A significant aspect of investigation in the range of medical imaging is early detection and categorization of brain tumours, which helps doctors choose the most practical way to proceed to rescue patient's lives [7].

For instance, doctors may employ radiation, surgery or chemotherapy to treat tumours. However, it always relies on the shape, nature, and size of a tumour. Clinically relevant

technology, such as Magnetic Resonance Images, produces extensive information on tumour and normal regions in the type of their slices. Furthermore, not all slices can be seen as tumours with the naked eye of a human. Therefore, correct assessment of a brain tumour necessitates the use of a skilled radiologist. Therefore, in order to identify a tumour in MR images without the use of humans, automated machines are always necessary [8]. The categorization of brain tumours has been approached from a variety of perspectives [9]. The investigators convey a variety of strategies, including super pixel-based brain tumour segmentation, multifractal features, salient structural topographies with RBF SVM kernel, and clustering-based segmentation with SVM[10]. ML methods rely on manually created characteristics, which limit the method's resilience. The efficiency of the deep learning-based algorithms, however, is substantially greater because they dynamically identify useful characteristics [1]. The current automated and semi-automated disorder analysis procedure's main goal is to create a reliable disease recognition system to help the physician with diagnosis and treatment planning [11]. The majority of image processing methods are utilized for tumour diagnosis. The goal of segmenting the image is to divide an image into uniform districts, thereby identifying the structures of the district [12]. The Magnetic Resonance Images plays an integral role in sophisticated research studies of the human brain. Magnetic Resonance scans may reveal important details about the composition of soft tissue. Additionally, the superiority of diagnosis and brain pathology are significantly enhanced by Magnetic Resonance imaging [3]. A doctor performs a biopsy to screen for tumours by taking a lesser sample of tissue and examining it under a microscope. Even though a biopsy can detect abnormalities with accuracy, people usually experience pain throughout the procedure. Second, doctors must be cognizant of the precise position and size of the tumour before doing surgery [13]. Therefore, MRI or CT scans are the most common approaches for evaluating the structure of a brain tumour. Nevertheless, MRI provides an accurate view of the structural development of brain tissues, whereas CT scans expose people to radiation that is harmful to their health [2].

Traditional ML strategies for categorization tend to concentrate on only limited or elevated features employ some manually created features to close these gaps, and call for effective feature extraction and categorization procedures. Deep CNNs, a recent innovation in DL, have attained achievement in the categorization of images [14]. Nowadays, a significant part of investigation in the health care zones is the automated recognition and segmenting the organs on medical images. There have been several techniques created that cover all geographical localizations and imaging modes. These researches, which make use of complicated blob-based algorithms, shared a recognition rates that seemed hard to enhance without complicating the feature extraction techniques significantly. Additionally, DL methods outperform typical machine learning techniques, which are constrained in their ability to analyse visual features in their natural form, time-consuming, dependent on professional expertise, and demanding a lot of work for parameter tuning [15].

In this study, three pathogenic categories of brain tumours precise and automatic categorization system including glioma, meningioma and pituitary tumour was provided. For extracting the features from brain MRI images, the remedy makes use of deep transfer learning model CNN [6]. It can be used in a variety of fields, such as object recognition, speech recognition, and image categorization. The DL mechanism is frequently utilized with CNN because it makes it simple to handle the hidden, input, and output layers [16]. Utilizing tested classifiers, the collected characteristics are categorised. Next, a thorough assessment of the suggested system is made. When tested on the open dataset, the suggested system outperformed all similar research in terms of classifier performance. Additionally, the suggested approach is found to deliver respectable results with fewer training data [6].

The following is a list of this article's main contributions:

- The system begins by collecting and processing a large set of input images using an MRI dataset.
- A Gabor filter is then employed to further analyze the input images that were generated.
- The impacted portion is then segmented through Otsu thresholding.
- A hybrid optimization technique combining the Bat and Grey Wolf performs feature extraction.
- After that, classification is carried out with the CNN-LSTM model.
- The efficiency of the developed approach is then verified and contrasted with those of other pre-trained models.

The residue of the article is arranged as follows: Section II deliberates the closely related studies. The proposed method, block diagram, flowchart and algorithms are thoroughly summarized in Section III. In Section IV, the experimental method of the suggested brain tumour categorization and recognition system is covered in detail. Section V illustrates the research findings and contrasts them with existing systems. The conclusion is organized in Section VI.

## II. RELATED WORK

Jaeyong Kang et al. [1] introduced a technique for categorising brain tumours using a combination of deep features and ML classifiers. The idea of transfer learning and various pre-trained DCNN are employed in this suggested framework to collect deep features from brain MRI. Different machine learning classifiers subsequently examine the deep features that were gathered. An ensemble of deep features consisting of the top 3 deep features that consistently outperform other machine learning classifiers is chosen, combined, and utilized to determine the result. Three distinct brain MRI datasets which are freely accessible are employed to compare the performance of machine learning classifiers, deep feature extractors, and an ensemble of deep features for the categorization of brain tumours. However in some circumstances, the Support Vector Machine with Radial Basis Function (RBF) kernel surpasses conventional ML classifiers, mainly for large datasets. According to experimental

observations, this method suggests that an ensemble of deep features can greatly increase efficiency. Even though our suggested strategy performs well, more research is required to minimize the model's structure so that it can be implemented on an actual medical diagnosis process utilizing knowledge filtering techniques. Siva Raja and Antony Viswasa rani [3] created a combined deep auto encoder with a Bayesian fuzzy clustering (BFC) method that is based on segmentation to categorize brain tumours. The non-local mean filter is originally utilized in the pre-processing phase for noise elimination purposes. Then, the BFC technique is employed to separate brain tumours. Following segmentation, reliable attributes are collected by utilizing Wavelet Packet Tsallis Entropy and Scattering Transform techniques. In order to categorise the tumour component for the brain tumour classification procedure, a combined strategy of the DAE oriented Jaya optimization algorithm combining softmax regression method is used. A MATLAB framework is applied to perform the suggested strategy. In comparison to other techniques, the simulated outcomes from the BRATS 2015 database demonstrated that the suggested strategy acquired a significant amount of classification accuracy of 98.5%. However, the more prevalent technique that will be employed to enhance the accuracy by combining more than one classifier based on the huge library of medical images and the precise classification strategy is not discussed.

Shahariar Alam et al. [2] presented a mechanism for identifying human brain tumours in an MRI image that combines the template-based K means and enhanced fuzzy C means (TKFCM) algorithms. Initially, the template-based K-means method is employed in this suggested technique to effectively choose a template created on the gray-level intensity of the image, which greatly initialises segmentation. Eventually, the enhanced FCM clustering method is employed for sensing tumour place by upgrading membership function that is derived on the basis of the characteristics of tumour images such as Energy, Homogeneity, Correlation, Dissimilarity, Entropy and Contrast. The revised membership is calculated by the cluster centroid distances to cluster data points by employing the Fuzzy C Means algorithm that gives better outcomes. According to simulated outcomes, the suggested technique is more effective at identifying diseased and normal brain tissues with only a slight loss in gray-level intensity. Additionally, compared to other techniques, this method predicts human brain tumours in a matter of seconds. However, the accuracy is left undiscussed. Rupa Ezhil Arasi and Suganthi [17] suggested a Soft Computing techniques of Clinical Support System for Classifying Brain Tumours. The brain MRI image is pre-processed by applying Genetic Optimized Median Filter in the presented Clinical Support System, and then the brain tumour zone is segmented by applying Hierarchical Fuzzy Clustering Technique. The GLCM feature extraction approach is utilized to capture the characteristics of the tumour region. The Brain Tumour Image Segmentation dataset is employed to accurately classify the tumour using the Lion Optimized Boosting Support Vector Machine method. As a result, the suggested clinical support system provides a comprehensive framework for the recognition and categorization of brain tumours, assisting the physicians in a proper assessment of the tumour. The findings

show that the suggested approach accurately categorises the tumour with a 97.69% accuracy rate. The primary benefit of the suggested approach is that it also evaluates tumour size and identifies the forms and phases of tumour. However, diagnostic errors are not recognized.

In order to improve images, Muhammad Sharif et al. [18] introduced a triangular fuzzy median filtering that supports in precise segmentation utilizing an unsupervised fuzzy system technique. In this method, Similar texture (ST) features are validated by utilising retrieved Gabor features over each person's tumours. Extreme learning machine ELM obtains these similar texture behaviours, and the reduction ELM omits one for tumour classification. On the BRATS 2013, 2012, 2015, 2014 datasets and on the 2013 Leader board, the method is tested. The suggested method yields superior outcomes and requires less computing time. However, managing distorted images and segmentation precision is not improved. Javaria Amin et al. [19] developed an innovative method on the basis of LSTM technique and MRI is offered to address the issues with automatic brain tumour categorization. To enhance the image quality of the multi-sequence MRI, N4ITK and 5×5 sized Gaussian filters are first utilised in this approach. The 4 layer deep LSTM framework is typically provided for classifying. The best hidden units including 200 HU and 225 HU are taken for every layer. In order to achieve superior outcomes, these disguised or hidden components were selected after intensive experimentation. The SISS-ISLES 2015 dataset and various BRATS dataset variants are employed to validate the findings. The approach was also tested using actual brain tumour patients from Pakistani ordinance factories, with a 0.97 DSC. The outcomes show that the proposed strategy gives radiologists additional assistance in accurately classifying brain tumours. The suggested technique had an accuracy rate of up to 98%. However, classifying subtumoural region and measuring the severity level of tumour region is not discussed.

Shah Rukh Khan et al. [13] introduced a Partial Tree, an association rule classifier with a sophisticated characteristic set to identify brain tumours according to their grade. The suggested method is evaluated by applying a 10-fold cross-validation procedure, and it is contrasted with other mechanisms including Random Forest, CART, Naïve Bayes and Random Tree. In this technique, threshold segmentation and masking are performed to magnetic resonance images as pre-processing processes prior to feature extraction. Depending on the grey level of the pixels, threshold segmentation isolates them into different sections. An intensity value called as the threshold determines categorization. The outcomes demonstrate that a partial tree with an enhanced feature set outperforms the existing approaches. Additionally, certain more sophisticated features should be utilized to boost performance. Arunkumar et al. [20] developed a new method of segmenting brain tissues from Magnetic resonance images. The main vision - based simulation methodologies utilized in the method are image segmentation, non-ROI filtering and image enhancement on the basis of texture and HOG features. ANN is employed in a fully automated framework technique for MRI brain tumour segmentation and categorizing to accurately pinpoint the

ROI's position. In order to evade non ROI and choose the exact object in brain MRI, the filtering out non ROI technique has been employed in perspective of histogram analysis. Nevertheless, using the textural features can determine the type of tumour. For the comparison of the automated and human segmentation processes, 200 MRI samples are employed. The results analysis proves that fully automatic model on the basis of trainable segmentation outperforms traditional methods and ROI texture-based brain diagnosis. 92.14% accuracy in diagnosis was recorded, with 94 specificity and 89 sensitivity. However, the brain trainable segmentation challenge was frustratingly difficult because of the enormous variability in brain tumour size and position in the images.

Bahadure et al. [21] introduced a comparative methodology on the basis of magnetic resonance images of brain tumour Segmentation and categorization by applying genetic algorithm. In this method, various segmentation strategies are compared to enhance the efficiency of tumour identification, and the better segmentation methodology is chosen by contrasting their segmentation scores. Additionally, the genetic algorithm is utilized for the automated identification of tumour stage in order to increase accuracy rate. Extraction of pertinent features and region computation assist the categorization phase choice. Depending on sensitivity, accuracy, specificity, segmentation score and dice similarity index coefficient, the study outcomes of the suggested methodology are assessed and verified for both quality and performance assessment on MR brain images. The experimental outcomes averaged a dice similarity index coefficient of 93.79%, indicating improved overlapping between radiologists' subjectively and automatically derived tumour regions. However, the discussion about a research of reliable technique for the huge medical image database and a discriminating classifier strategy by integrating more than one classifier is not recognized. Diaz-Pernes et al. [22] introduced a fully automatic brain tumour segmentation and classification technique that employs use of a multiscale DCNN. The functioning of the human visual system aided as motivation for this approach. The proposed neural framework is capable of performing tumor-specific MRI image analysis. The technique's efficiency on a dataset of 3064 MRI image slices from 233 patients that is publicly visible is contrasted with other traditional ML and DL approaches. In the assessment, this approach significantly performed with a tumour classification accuracy of 0.973. This technique helps doctors to diagnose brain tumours, and the presented segmentation and categorization technique can be utilised to solve various imaging issues in the field of medicine. However, it is not explained how the suggested multiscale CNN for segmentation could be applied to other research areas, like satellite imaging.

The aforementioned literature review provides a clear picture of the techniques that have been developed, the more common technique that will be used to increase accuracy by combining multiple classifiers based on the vast library of medical images, and the precise classification approach are not included, and diagnostic errors that are not recognized. There is no discussion of classifying the sub tumoural region

or determining the tumor location's severity level. The huge variation in size and location of the brain tumors in the images made the brain trainable segmentation problem painfully challenging. To solve these problems, the BGWO-CNN-LSTM approach is suggested in this study for the detection and classification of brain tumors using MR images. The observations support the assertion that the suggested strategy makes it simple for clinical experts to make decisions about diagnosis and scanning.

### III. PROPOSED METHODOLOGY

The presentation is the automatic process for the lesion and imaging stages of brain tumour identification. MRI tests are done on the proposed system. As a result of its high contrast, spatial resolution, and low radiation MRI seems to be more effective at detecting tumours. The location and size of a brain tumour can be determined by MR scans [12]. The presented approach contains five main steps: image acquisition, pre-processing, segmentation of images, feature extraction and classification. Various methodologies are used during pre-processing to split the region of interest. The Gabor filter is utilized for the pre-processing stage. The next phase is image segmentation, where the normal and abnormal regions are separated using K means clustering. Next, the feature extraction is done by applying Bat and Grey Wolf Optimization (BGWO) algorithm. Finally, CNN-LSTM is utilized for classification. The developed method is outlined in Fig. 1.

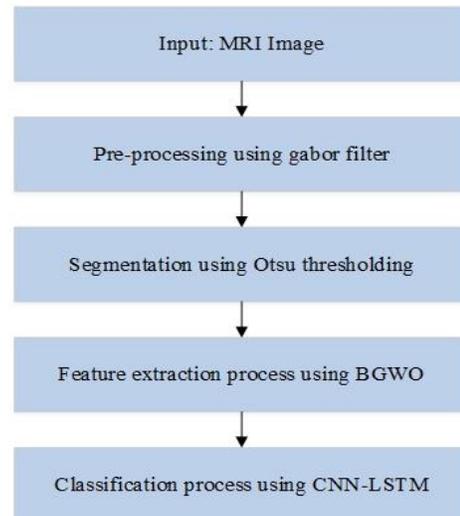


Fig. 1. Overall process of the developed method.

#### A. Data Collection

The effectiveness of the developed detection system relies on the database that is utilized in every medical assessment operation. The dataset utilised in this research was obtained from BRATS 2015 that contains a distinct version of BRATS that is designed to address medical imaging challenges. The BRATS 2015 imaging dataset, that can display an upgraded form of new cases reported with greater efficiency, was derived from BRATS 2012 and BRATS 2013 [16]. The dataset contains 6100 MRI images of the human brain that are categorized into 4 categories: meningioma, pituitary, glioma,

and normal. The dataset is splitted into 80% of training data and 20% of testing data. The description of database is shown in Table I.

TABLE I. DATASET DESCRIPTION

Disease type	Training data	Testing data	Total
Pituitary	1205	195	1400
Meningioma	1327	473	1800
Glioma	534	366	900
Normal	1602	398	2000

### B. Pre-processing

The brain MRI datasets entirely contain unwanted gaps and areas, which leads to inadequate categorization accuracy. Therefore, it is vital for image cropping to eliminate unnecessary portions and use relevant data. The cropping approach is applied for calculating extreme points. For pre-processing, import the primary MR images. Then, in order to create binary images, perform thresholding to the MR images. In this research, Gabor filter is applied. When the Gabor filter is utilized for texture features, it effectively examines if the image contains any particular frequency information or particular directions in a restricted region around the pixel or region of evaluation. Gabor filters are utilized in the study of multi-resolution images which is like human vision cortex since every filter exhibits a sensory neuron which is responsive to a certain frequency. This is because of the band-pass nature and chosen direction qualities of Gabor filters [18].

Gaussian function in the  $f(a, b)$  domain's spatial coordination is defined using the Gabor filter. Assume the Fourier transform of  $f(a, b)$  denoted by  $F(m, n)$  that is a function of the frequency components  $(m, n)$  given in Eq. (1).

$$f(a, b, \sigma, \beta) = \left( \frac{1}{2\pi\sigma_a\sigma_b} \right) \exp \left[ -\frac{1}{2} \left\{ \frac{a^2}{\sigma^2} \right\} + 2\pi r\beta a \right] \quad (1)$$

In the Gabor filter, the Gaussian window is adjusted by  $\sigma$  among the appropriate axis while the Gaussian function is combined in the Fourier domain.  $\beta$  represents the Gabor filters' centre frequency. The filter results are displayed in Eq. (2) with a good response at the frequency centre.

$$K(\mu, w, \sigma, \beta) = \exp \left\{ -\frac{1}{2} \left( \frac{(\mu-\beta)^2}{\sigma_\mu^2} + \frac{w^2}{\sigma_w^2} \right) \right\} \quad (2)$$

$$\text{Where } \sigma_\mu = \frac{1}{2\pi\sigma_a} \text{ and } \sigma_w = \frac{1}{2\pi\sigma_b}$$

### C. Segmentation using Otsu Thresholding

After pre-processing, Otsu thresholding is utilized for segmentation process. Otsu is an automated threshold selection technique for segmenting data depending on region. Gray levels are used in the unsupervised, nonparametric Otsu threshold method. Otsu threshold criterion utilises an image's gray-level histogram, and the threshold procedure determines a normalised value in the [0, 1] range [23]. The probability distribution is represented in Eq. (3).

$$a(t) = \frac{ht}{H} \quad (3)$$

Where,  $a(t)$  is the probability distribution,  $H$  represents the total no. of image pixels, and  $h$  is the histogram count for pixel value  $t$ .

Eq. (4) and (5) employ the probability distribution  $a(t)$  to derive the class probability. Every image pixel is divided into the classes of background and object, with a threshold, separating them. The class probability is expressed as,

$$u_0 = \sum_{m=1}^l a(t) \quad (4)$$

$$u_1 = \sum_{m=l+1}^k a(t) \quad (5)$$

Following, the class mean which is denoted by  $\mu$  is represented as,

$$\mu_0 = \sum_{m=1}^l \frac{ta_t}{u_0} \quad (6)$$

$$\mu_1 = \sum_{m=l+1}^k \frac{ta_t}{u_1} \quad (7)$$

Eq. (6) and (7) are substituted in Eqns. (8) and (9), respectively.

The following equations are employed to form the class variance equation:

$$\sigma_0 = \sum_{m=1}^l [t - \mu_0] 2 \frac{a(t)}{u_0} \quad (8)$$

$$\sigma_1 = \sum_{m=l+1}^k [t - \mu_1] 2 \frac{a(t)}{u_1} \quad (9)$$

The formula for weighted within-class variance is illustrated in Eq. (10)

$$\sigma_u^2 = u_0\sigma_0^2 + u_1\sigma_1^2 \quad (10)$$

Where,  $u_0$  is the weight of the background,  $u_1$  is the weight of foreground,  $\sigma_0$  is the variance of background and  $\sigma_1$  is the variance of the foreground.

The performance of segmenting the brain tumour is evaluated through the performance rate after the Otsu approach is applied to the synthetic image which is given in Eq. (11).

$$\text{Performance} = \frac{\text{No. of pixels of object after segmentation}}{\text{No. of pixels of object before segmentation}} \quad (11)$$

Generally, a successful segmentation of an image occurs when all of the object's pixels are separated from the background without any additional or subtracted pixels. When segmenting an image, the performance is less than 1, which suggests that some pixels related to the object were mistakenly categorised as background and object. If the rate of performance is 1, then an object's pixels have all been completely segmented.

### D. Feature Extraction and Selection using BGWO Algorithm

1) *Bat optimization algorithm*: A meta-heuristic search technique called the BAT algorithm [24] imitates the action of bats. This optimization method utilizes the echolocation system of bats to find food and distinguish it from other objects. Bats make quick, loud pulses that detect an echo and return to their ears. BAT determines the distance, kind of

object and the time it takes for it to return. The bat's position and velocity matrices,  $a$  and  $p$ , are updated by bat algorithm in the  $d$ -dimensional search area is represented in Eq. (12) [25].

$$a_m^l = a_m^{l-1} + (p_m^{l-1} - p_{best}) \times g_m \quad (12)$$

Where  $l$  is the current iteration, and  $p_{best}$  is the best global solution.

2) *Grey wolf optimization algorithm*: The grey wolf optimizer (GWO) method [26] is a swarm-based method that derives from nature and imitates the hierarchical society of wolves and their behaviour in encircling, approaching and attacking the prey. The grey wolves social behaviour during the hunting process is portrayed statistically to solve an optimization issues by applying the GWO algorithm. The wolves in GWO iterations assess the potential hunting circumstances and revise their status as necessary. The encircling process' mathematical expression is given by,

$$\vec{G}(s+1) = \vec{G}_a(s) - \vec{P} \times \vec{R} \times \vec{G}_a(s) - \vec{G}(s) \quad (13)$$

In the equation preceding,  $s$  stands for the current iteration,  $\vec{G}_a$  and  $\vec{G}$  stand for the hunt and hunter position vectors and  $\vec{P} = 2\vec{p} \cdot \vec{c}_1 - \vec{p}$  and  $\vec{R} = 2\vec{c}_2$  stands for coefficient vectors.  $\vec{c}_1$  and  $\vec{c}_2$  are random numbers between 0, 1 that permit the wolves change their position in the hunt space. The best exploration factor equation is represented in Eq. (14)

$$\vec{G}(s+1) = \frac{\vec{G} + \vec{G}_2 + \vec{G}_3}{3} \quad (14)$$

A novel hybrid combination of BGWO using the fitness values is given in Eqn. (15),

$$a_m^l = a_m^{l-1} + (p_m^{l-1} - p_{best}) \times \vec{G}(s+1) \quad (15)$$

### E. Classification

1) *Convolutional neural network*: The CNN is the deep neural network. It aims to discover the underlying and intrinsic characteristics from guided processing of 2-Dimensional or 3-Dimensional images. These characteristics are suitable for classifying anatomical structures and identifying aberrant structures. An input layer is associated with a number of pooling layers, output layer and convolutional layer in a standard CNN architecture.

a) *Convolutional layer*: The convolutional layers perform a convolution operation by using convolution kernels and the raw input data to create new attribute values. The model was developed to collect features from dataset images, so the input data should be in the procedure of structured matrix. When compared to the input matrix, the convolution kernel is comprehended as a narrow window that organises coefficient values into a matrix. This window "slides" around the input matrix, performing a convolution process on every

patch while moving. A convolved structure, is a feature variable established by the coefficient values and the allocated dimension element of the filter. Convolved features which are often more useful than the principal features of the input data can be produced by applying various convolution kernels to the input data. Hence, the approach performs better. The basis of a CNN is a convolutional layer, because the majority of computations are completed at this layer. It is a quality extraction layer that pulls out the regional features via the filters and produces a convolutional calculated feature map and exits the kernel function and goes to the pooling layer. The convolutional layer is expressed in Eq. (16).

$$P_m^{(a)} = \sigma(G_m^{(a)} + \sum_{n=1}^{s(a-1)} P_n^{(a-1)} * U_{m,n}^{(a)}) \quad (16)$$

Where the operator  $*$  denotes the convolution operation,  $\sigma$  is an activation matrix, and  $U_{m,n}^{(a)}$  is the filter linking the  $n^{th}$  feature map in layer  $a-1$  with the  $m^{th}$  feature map in layer  $a$  is a function that is employed to increase nonlinearity.

b) *Pooling layer*: Typically, the pooling layer is applied following the convolutional layer. The pooling layer's job is to streamline the data in the output of the layer of convolution. The pooling layer creates a compressed feature map using each feature map's data from the convolutional layer. The most popular techniques are max-pooling and average pooling. There is no learning happening in this tier. Size  $N \times N$  filters have been chosen in this layer. The average pooling and max pooling layer is represented in Eq. (17) and (18).

$$\bar{a} = \frac{1}{L} \sum_{(m,n) \in G} a_{m,n} \quad (17)$$

$$a_{max} = \max_{(m,n) \in G} (a_{m,n}) \quad (18)$$

Where  $a_{m,n}$  is the number of each pixel in area  $G$  and  $L$  is the area's pixel count.

c) *Dense layer*: In Dense layer, the Long-Short Term Memory (LSTM) method has been utilized. In particular, LSTM neural networks [27] are a subclass of recurrent neural networks with learning capabilities across time employing feedback connections. This technique develops short-term memory and gathers data from it by utilizing cyclic links on their hidden layer and collect information from time series and sequences. A memory cell and the three major gates of input forget and output make up every LSTM unit. By utilizing this framework, the LSTM choose which information needs to be "forgotten" and which needs to be "remembered," creating a controlled data flow and learning for the long term dependencies. Eq. (19) predicts the performance operation of LSTM unit. The layers of the developed method are shown in Fig. 2.

$$a_u = \sigma(P_u m_s + R_u n_{s-1} + t_u) \quad (19)$$

Where  $R$  and  $P$  are weight matrices,  $m_s$  represents the input,  $\sigma$  is the sigmoid function, and  $t$  is the bias term vector.

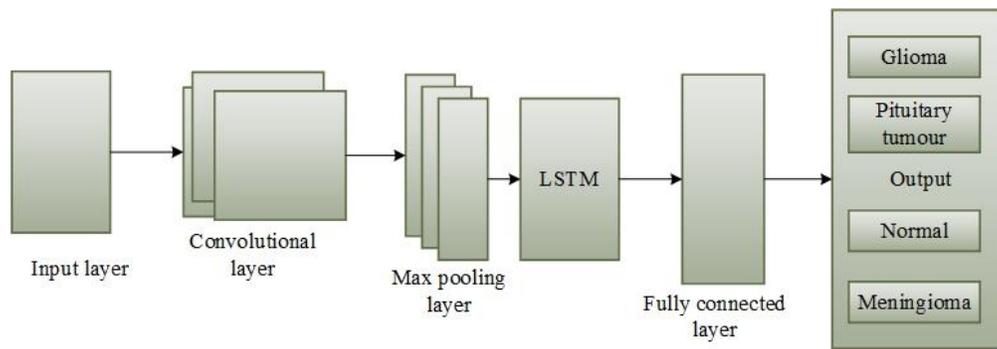


Fig. 2. Layers of the developed CNN and LSTM model.

d) *Output layer*: The output layer also known as fully connected layer is the completely linked layer's neurons rely on all regions of the brain's prior layer. In this layer, data is transformed into a 1-D matrix beneath the layer. There may be variations in each model's overall amount of fully connected layers. Eq. (20) is employed for feed forward in this layer.

$$a_m^s = \sum_n u_f^{s-1} v_n^{s-1} \quad (20)$$

Where  $s$  is the number of layer,  $v_n^s$  is the value in the created output layer,  $m$  and  $n$  is the number of neuron,  $u_f^{s-1}$  is

the hidden layer weight,  $v_n^{s-1}$  is the input neuron input and  $a_m^s$  is the value of the activation function in the output layer.

The brain tumours have been correctly classified based on MRI images. Pituitary tumour, glioma, normal brain and meningioma have been discovered. The CNN-LSTM framework utilizes Bat and Grey Wolf Optimization (BGWO) algorithm for extracting the features and the overall process BGWO-CNN-LSTM is shown in Algorithm 1 and Fig. 3.

---

**Algorithm 1:** BGWO-CNN-LSTM mechanism

---

**Input:** Magnetic Resonance Images

**Output:** Pituitary, Meningioma, Glioma, and Normal

Import input image data

Let  $I$  be the input data that is taken for analysis

$$I = \{I_1, I_2, I_3 \dots\}$$

Pre-processing of images

Segmentation of images

Feature extraction

Initialize the bat's population randomly

while ( $s' < \text{Maximum number of iterations}$ )

Adjust frequency and generate new position

Update velocity and position using eqn. (14)

else go to next step

if ( $\text{random} > P_m$ )

Select an image randomly among the best positions

Calculate local position among the selected best position

Else

calculate the global best position

Calculate the fitness of every search agent

Update local and global best positions

$$s' = s' + 1$$

end if

Classification

Classifying as Pituitary, Meningioma, Glioma, and Normal

end if

end while

end

---

//Gabor filter

//Otsu thresholding

//Bat Grey Wolf Optimization

//CNN-LSTM classifier

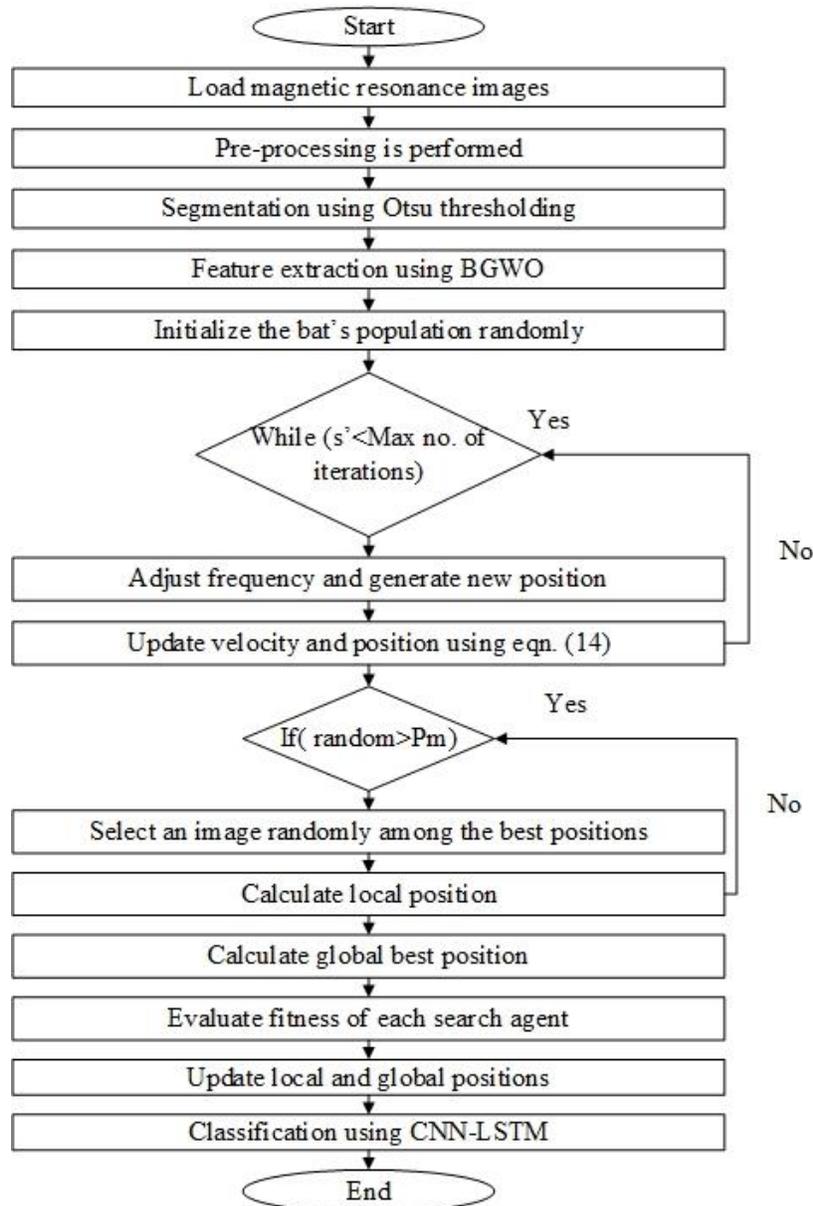


Fig. 3. BGWO-CNN-LSTM workflow diagram.

#### IV. RESULTS

The developed method is examined using MRI dataset. Using the Gabor filter, 6100 MRI scans of brain tumours were gathered and pre-processed. The tumours in the images are so severe. So, it is impossible for the average person to quickly spot them. Pre-processing MR images is crucial for improving the vision impact of the image before processing. Typically, the dataset's gathered images are of bad quality and the noise should be filtered out and the image is sharpened. Gabor filter is utilised as a pre-processing step. Following that many traits are initially implicitly collected. With the aid of an enhanced Otsu thresholding method, the tumour is then identified. After that, the feature extraction method utilizes the bat and grey wolf algorithm. Subsequently CNN-LSTM is utilised for classification. The proposed model achieves performance matrices of recall, accuracy, and specificity. 4100 images of

brain tumours and 2000 photos of tissue are found by extracting and segmenting image features. Utilizing efficient spot images, training and testing dataset is produced. The proposed model's effectiveness is shown, and it achieves the best levels of recall, specificity, and accuracy in the identification of brain tumours. In every MRI imaging of a brain tumour, there is an error rate based on abnormal tissue. Depending on the true negative, true positive, false positive and false negative values, these can be quantified. The recall, accuracy and specificity of the technique have been evaluated on all of the images in the dataset for this research.

##### A. Accuracy

One of the frequently employed metrics for classification techniques is accuracy. It indicates the proportion of accurate estimates of overall predictions. The evaluation of actual

classification is known as accuracy. According to image analysis, the accuracy is a per cent that represents the total amount of pixels that have been correctly classified in relation to the total amount of pixels in the image. It assesses every single correctly placed pixel in an image. Accuracy is expressed in Eq. (21).

$$Accuracy = \frac{True_{pos} + True_{neg}}{True_{pos} + True_{neg} + False_{pos} + False_{neg}} \quad (21)$$

### B. Specificity

The quantity of precisely determined true negatives is measured by specificity. Using Eq. (22), the specificity value is calculated as,

$$Specificity = \frac{True_{neg}}{True_{neg} + False_{pos}} \quad (22)$$

### C. Recall

Recall is the ratio of true positives and false negatives to correct positive forecasts. The percentage of forecasts that have been appropriately identified as tumour is expressed. Eq. (23) is employed to represent recall.

$$Recall = \frac{True_{pos}}{True_{pos} + False_{neg}} \quad (23)$$

TABLE II. PERFORMANCE MATRIX COMPARISON ON THE BASIS OF ACCURACY, SPECIFICITY AND RECALL

Method	EKF-SVM	Genetic Algorithm	ANN	Proposed BGWO-CNN-LSTM
Accuracy	98.02%	92.03%	92.14%	99.74%
Specificity	94.15%	91.42%	94%	99.54%
Recall	96.44%	92.36%	89%	99.23%

The test results of the BGWO-CNN-LSTM classifier with those of other classifier methods including Extended Kalman filter with Support Vector Machine (EKF-SVM) [28], Genetic algorithm [21], Artificial Neural Networks (ANN) [20] based on the statistical characteristics of recall, accuracy and specificity is illustrated in Table II. On the basis of disease categories such as glioma, pituitary tumour, meningioma and normal, the findings are compared. 80% of image dataset are employed for training and 20% of images are employed for testing the data. The performance matrix of the suggested approach's recall, specificity and accuracy were found to be 99.23%, 99.54% and 99.74% which is higher than the existing approaches of EKF-SVM, Genetic algorithm and ANN is illustrated in Fig. 4.

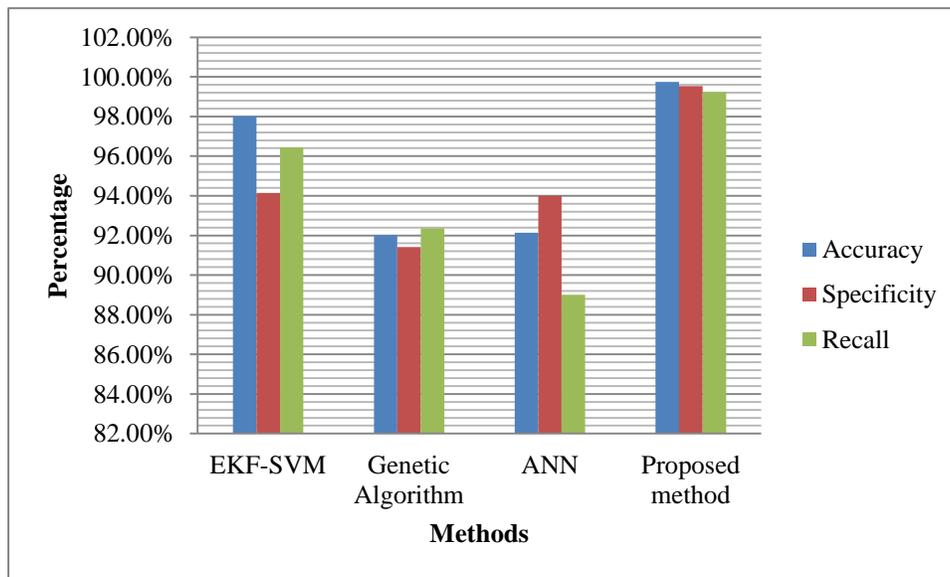


Fig. 4. Performance comparison of the developed and existing models.

## V. DISCUSSION

The brain MRI dataset was used in this research work to apply a combined mechanism to recognize and classify the tumor on the images. The created method uses supervised hybrid CNN and LSTM algorithms and is intended to discriminate between normal and pathological tumors in brain pictures. Brain tumour detection techniques including Extended Kalman filter with Support Vector Machine, Genetic algorithm and ANN are trained and evaluated in order to compare their characteristics with those of traditional networks. The findings demonstrate that the suggested BGWO-CNN-LSTM model, out of these four strategies, produces greater specificity, accuracy and recall as shown in Fig. 4. A degree of recall, specificity and accuracy of 99.23%,

99.54% and 99.74% are attained using the BGWO-CNN-LSTM method. The comparison demonstrates that the suggested approach outperformed the alternatives. The suggested approach demonstrates that the BGWO-CNN-LSTM is the better approach for the recognition and classification of a brain tumour.

## VI. CONCLUSION

One of the serious disorders is brain tumour identification and classification due to aberrant cell proliferation or portable spread across the body. This research work has applied a combined mechanism on brain MRI images to identify and categorise the tumour utilising the MRI dataset. Employing supervised hybrid CNN and LSTM approaches, the developed

method is designed to distinguish among normal and abnormal tumours in brain images. The input images have undergone the main pre-processing processes of normalisation, as well as the extraction of major characteristics from the pre-processed image using the Gabor filter and threshold-based segmentation approach called Otsu thresholding. To categorize brain MRI images, hybrid CNN and LSTM algorithms are applied to the labelled segmented features. It is utilized to categorize the tumours such as pituitary tumour, glioma, meningioma and normal brain. Finally, the proposed approach achieved 99.74% accuracy, 99.23% recall and 99.54% specificity. Comparison of supervised and unsupervised learning in recent technologies verifies that the suggested BGWO-CNN-LSTM method works better than other well-known CNN-based architectures for classifying the tumours. Further research could improve feature extraction algorithms by incorporating additional texture and form features and employing big datasets.

#### REFERENCES

- [1] J. Kang, Z. Ullah, and J. Gwak, "MRI-Based Brain Tumor Classification Using Ensemble of Deep Features and Machine Learning Classifiers," *Sensors*, vol. 21, no. 6, p. 2222, Mar. 2021, doi: 10.3390/s21062222.
- [2] M. S. Alam et al., "Automatic Human Brain Tumor Detection in MRI Image Using Template-Based K Means and Improved Fuzzy C Means Clustering Algorithm," *Big Data Cogn. Comput.*, vol. 3, no. 2, p. 27, May 2019, doi: 10.3390/bdcc3020027.
- [3] P. M. Siva Raja and A. V. rani, "Brain tumor classification using a hybrid deep autoencoder with Bayesian fuzzy clustering-based segmentation approach," *Biocybern. Biomed. Eng.*, vol. 40, no. 1, pp. 440–453, Jan. 2020, doi: 10.1016/j.bbe.2020.01.006.
- [4] M. M. Badža and M. Č. Barjaktarović, "Classification of Brain Tumors from MRI Images Using a Convolutional Neural Network," *Appl. Sci.*, vol. 10, no. 6, p. 1999, Mar. 2020, doi: 10.3390/app10061999.
- [5] M. Sharif, J. Amin, M. Raza, M. Yasmin, and S. C. Satapathy, "An integrated design of particle swarm optimization (PSO) with fusion of features for detection of brain tumor," *Pattern Recognit. Lett.*, vol. 129, pp. 150–157, Jan. 2020, doi: 10.1016/j.patrec.2019.11.017.
- [6] S. Deepak and P. M. Ameer, "Brain tumor classification using deep CNN features via transfer learning," *Comput. Biol. Med.*, vol. 111, p. 103345, Aug. 2019, doi: 10.1016/j.compbiomed.2019.103345.
- [7] E. Irmak, "Multi-Classification of Brain Tumor MRI Images Using Deep Convolutional Neural Network with Fully Optimized Framework," *Iran. J. Sci. Technol. Trans. Electr. Eng.*, vol. 45, no. 3, pp. 1015–1036, Sep. 2021, doi: 10.1007/s40998-021-00426-9.
- [8] A. Rehman, M. A. Khan, T. Saba, Z. Mehmood, U. Tariq, and N. Ayesha, "Microscopic brain tumor detection and classification using 3D CNN and feature selection architecture," *Microsc. Res. Tech.*, vol. 84, no. 1, pp. 133–149, Jan. 2021, doi: 10.1002/jemt.23597.
- [9] M. Toğaçar, B. Ergen, and Z. Cömert, "BrainMRNet: Brain tumor detection using magnetic resonance images with a novel convolutional neural network model," *Med. Hypotheses*, vol. 134, p. 109531, Jan. 2020, doi: 10.1016/j.mehy.2019.109531.
- [10] M. A. Khan et al., "Brain tumor detection and classification: A framework of marker-based watershed algorithm and multilevel priority features selection," *Microsc. Res. Tech.*, vol. 82, no. 6, pp. 909–922, Jun. 2019, doi: 10.1002/jemt.23238.
- [11] V. Rajinikanth, A. N. Joseph Raj, K. P. Thanaraj, and G. R. Naik, "A Customized VGG19 Network with Concatenation of Deep and Handcrafted Features for Brain Tumor Detection," *Appl. Sci.*, vol. 10, no. 10, p. 3429, May 2020, doi: 10.3390/app10103429.
- [12] J. Amin, M. Sharif, M. Yasmin, and S. L. Fernandes, "A distinctive approach in brain tumor detection and classification using MRI," *Pattern Recognit. Lett.*, vol. 139, pp. 118–127, Nov. 2020, doi: 10.1016/j.patrec.2017.10.036.
- [13] S. R. Khan, M. Sikandar, A. Almogren, I. Ud Din, A. Guerrieri, and G. Fortino, "IoT-based computational approach for detecting brain tumor," *Future Gener. Comput. Syst.*, vol. 109, pp. 360–367, Aug. 2020, doi: 10.1016/j.future.2020.03.054.
- [14] Z. N. K. Swati et al., "Brain tumor classification for MR images using transfer learning and fine-tuning," *Comput. Med. Imaging Graph.*, vol. 75, pp. 34–46, Jul. 2019, doi: 10.1016/j.compmedimag.2019.05.001.
- [15] S. A. Abdelaziz Ismael, A. Mohammed, and H. Hefny, "An enhanced deep learning approach for brain cancer MRI images classification using residual networks," *Artif. Intell. Med.*, vol. 102, p. 101779, Jan. 2020, doi: 10.1016/j.artmed.2019.101779.
- [16] M. O. Khairandish, M. Sharma, V. Jain, J. M. Chatterjee, and N. Z. Jhanjhi, "A Hybrid CNN-SVM Threshold Segmentation Approach for Tumor Detection and Classification of MRI Brain Images," *IRBM*, vol. 43, no. 4, pp. 290–299, Aug. 2022, doi: 10.1016/j.irbm.2021.06.003.
- [17] P. R. E. Arasi and M. Suganthi, "A Clinical Support System for Brain Tumor Classification Using Soft Computing Techniques," *J. Med. Syst.*, vol. 43, no. 5, p. 144, May 2019, doi: 10.1007/s10916-019-1266-9.
- [18] M. Sharif, J. Amin, M. Raza, M. A. Anjum, H. Afzal, and S. A. Shad, "Brain tumor detection based on extreme learning," *Neural Comput. Appl.*, vol. 32, no. 20, pp. 15975–15987, Oct. 2020, doi: 10.1007/s00521-019-04679-8.
- [19] J. Amin, M. Sharif, M. Raza, T. Saba, R. Sial, and S. A. Shad, "Brain tumor detection: a long short-term memory (LSTM)-based learning model," *Neural Comput. Appl.*, vol. 32, no. 20, pp. 15965–15973, Oct. 2020, doi: 10.1007/s00521-019-04650-7.
- [20] N. Arunkumar, M. A. Mohammed, S. A. Mostafa, D. A. Ibrahim, J. J. P. C. Rodrigues, and V. H. C. Albuquerque, "Fully automatic model-based segmentation and classification approach for MRI brain tumor using artificial neural networks," *Concurr. Comput. Pract. Exp.*, vol. 32, no. 1, Jan. 2020, doi: 10.1002/cpe.4962.
- [21] N. B. Bahadure, A. K. Ray, and H. P. Thethi, "Comparative Approach of MRI-Based Brain Tumor Segmentation and Classification Using Genetic Algorithm," *J. Digit. Imaging*, vol. 31, no. 4, pp. 477–489, Aug. 2018, doi: 10.1007/s10278-018-0050-6.
- [22] F. J. Díaz-Pernas, M. Martínez-Zarzuela, M. Antón-Rodríguez, and D. González-Ortega, "A Deep Learning Approach for Brain Tumor Classification and Segmentation Using a Multiscale Convolutional Neural Network," *Healthcare*, vol. 9, no. 2, p. 153, Feb. 2021, doi: 10.3390/healthcare9020153.
- [23] Z. Y. Tan, S. N. Basah, H. Yazid, and M. J. A. Safar, "Performance analysis of Otsu thresholding for sign language segmentation," *Multimed. Tools Appl.*, vol. 80, no. 14, pp. 21499–21520, Jun. 2021, doi: 10.1007/s11042-021-10688-4.
- [24] V. Sathananthavathi and G. Indumathi, "BAT algorithm inspired retinal blood vessel segmentation," *IET Image Process.*, vol. 12, no. 11, pp. 2075–2083, Nov. 2018, doi: 10.1049/iet-ipr.2017.1266.
- [25] T.-T. Nguyen, J.-S. Pan, and T.-K. Dao, "A Compact Bat Algorithm for Unequal Clustering in Wireless Sensor Networks," *Appl. Sci.*, vol. 9, no. 10, p. 1973, May 2019, doi: 10.3390/app9101973.
- [26] B. Mohammadi, Y. Guan, P. Aghelpour, S. Emamgholizadeh, R. Pillco Zolá, and D. Zhang, "Simulation of Titicaca Lake Water Level Fluctuations Using Hybrid Machine Learning Technique Integrated with Grey Wolf Optimizer Algorithm," *Water*, vol. 12, no. 11, p. 3015, Oct. 2020, doi: 10.3390/w12113015.
- [27] I. E. Livieris, E. Pintelas, and P. Pintelas, "A CNN-LSTM model for gold price time-series forecasting," *Neural Comput. Appl.*, vol. 32, no. 23, pp. 17351–17360, Dec. 2020, doi: 10.1007/s00521-020-04867-x.
- [28] B. Chen, L. Zhang, H. Chen, K. Liang, and X. Chen, "A novel extended Kalman filter with support vector machine based method for the automatic diagnosis and segmentation of brain tumors," *Comput. Methods Programs Biomed.*, vol. 200, p. 105797, Mar. 2021, doi: 10.1016/j.cmpb.2020.105797.

# Multi Feature DCR based Drug Compound Selection and Recommendation System for Efficient Decision-Making using Genetic Algorithm

ST. Aarthy<sup>1</sup>, Dr. J. L. Mazher Iqbal<sup>2</sup>

Research Scholar, Dept. of Electronics and Communication Engineering<sup>1</sup>

Vel Tech Rangarajan Dr. Sagunthala R & D Institute of Science and Technology, Avadi, Chennai, Tamil Nadu, India<sup>1</sup>

Professor, Dept. of Electronics and Communication Engineering<sup>2</sup>

Vel Tech Rangarajan Dr. Sagunthala R & D Institute of Science and Technology, Avadi, Chennai, Tamil Nadu, India<sup>2</sup>

**Abstract**—The performance of treating the cardiac diseases is dependent on the kind of drug being selected. There exist numerous decisive support systems which work according to certain characteristics and factors like drug availability, and popularity. Still, they struggle to achieve expected performance in supporting the medical practitioner. To handle this issue, a multi feature drug curing rate based drug compound selection and recommendation system (MDCRSR) is presented. The method utilizes medical histories and data set of various medical organization around the disease considered. Using the traces, the method identifies the drug compounds and features to perform preprocessing which eliminates the noisy data points. Further, the features of the traces are extracted to perform training with genetic algorithm. At the test phase, the method estimates the fitness measure for different drug combination and compounds by measuring their Drug Curing Rate (DCR). The method performs cross over and mutation to produce various populations of drug compounds. According to the curing rate, the drug compound pattern or population is selected and ranked. The ranked results are populated to the medical practitioner. The method improves the performance of recommendation system as well as drug compound selection.

**Keywords**—Decisive support Systems; GA; DCR; drug selection; compound selection; fitness; recommendation system; cardiac disease; RMDCRSR

## I. INTRODUCTION

The human society faces variety of diseases but not all of them are harmful and claim the human life, but there are few diseases which would claim the human life without giving any time. The cardiac disease is one among them which occurs in various ways and they can be classified as Dextrocardia, Tachycardia, Bradycardia, Hyperkalcaemia, Sinoatrial block, and Myocardialischaemia. Among these diseases, few of them instantly block the heart and leads to cardiac arrest and lead to death. However, by identifying the disease at the early stage, they can be treated with set of drugs. For other diseases, there are varieties of drugs available in various compounds, each of them differs with their efficiency in curing the disease. Here it is about the selection of exact drug for any disease with the available drug compounds.

The medical practitioners are capable of identifying the disease and even they require set of automated decisive

support system in the classification and detection of any disease. Also, they have lot of drugs and compounds in front of them, and they can choose any of them to serve with patients. But, the efficiency of curing the disease cannot be justified and the performance of curing the disease is highly dependent on the combination of drugs being selected. As, there are numerous drugs, the medical practitioner would confuse in the selection of exact drug and compounds. To handle this issue, a real time drug and compound selection approach is discussed.

The recommendation systems are over the decisive support systems which consider variety of medical logs and features. According to the features and traces, the recommendation system identifies set of features in terms of drugs and would measure their performance according to their success and failure. By measuring the value of performance, they can be ranked and populated to the medical practitioner to serve the patient. For example, towards cardiac disease there are many drugs available like Aspirin, Lidocaine, Disopyramide, Procainamide which are used towards arrhythmic diseases. However, these drugs are available, the performance of the drug must be considered while selecting a drug.

The optimal selection of drug is dependent on how efficient the drug compound is in curing the disease. It is more important to identify the exact drug and to perform this, the Genetic algorithm has been adapted in this paper. Genetic algorithm is a scientific approach of searching the drug more effective in curing the disease. The genetic algorithm is applied in several scientific problems of medical issues and the same can be used in the selection of drugs in this model. The drugs of any disease would come on different milligrams and applied to several diseases. The GA algorithm would search on the optimal combination of drug towards any disease by measuring Drug Curing Rate (DCR). The RMDCRSR algorithm estimates the fitness.

## II. RELATED WORKS

There are number of approaches available towards recommendation generation and drug selection for heart diseases. This section details set of approaches related to the problem.

A sentiment analysis-based drug recommendation scheme is presented in [1], which take the reviews of sentiment analysis and support the decision-making problem. The sentiment measurement scheme uses the drug rating and selects according to the patient conditions. Accordingly, recommendations are generated. Similarly, in [2], an implicit feedback based approach cross recommendations (IFCR) which uses epileptics' medical history in identifying the relation of syndrome among drugs. A context aware approach towards hypertensive drugs is presented to support recommendations on personalized scope [3]. The method uses Semantic Web Rule Language (SWRL) to generate recommendations. A detailed review on several recommendation system is presented in [4], to support healthcare professionals.

A user preference-based recommendation scheme is developed in [5], where the recommendations are generated with two approaches and the user can select accordingly. An android based application is developed to support the selection of required medication to manage diabetic in [6]. The recommendations are generated according to the input made.

A group-based approach is discussed in [7], with a consensus reaching process to stimulate the recommendations related to the group users. An evidence-based recommendation scheme is presented in [8], which uses physical activity, exercise in diabetes patients. A naïve bayes classification approach towards disease prediction and drug recommendation is presented in [9], which uses the profile of patients like blood pressure, heart rate, cold and fever in classification where the drug selection is performed according to the symptoms. Similarly, for the hypertensive patients a physical activity recommendation system is presented in [10], which uses the user profile in recommendation generation named HyperModel2PAR.

A genetic algorithm based multi expert scheme (GA) MES-GRS-GA is presented in [11], which discuss consensus scheme towards group recommendation system. A predictive system is designed towards supporting cardiac disease in [12], which analyzes various machine learning algorithms and their results. A Fourier transformation-based heart disease prediction system (FTHDPS) is presented towards predicting chronic heart diseases with time series dat. ANN has been used in recommendation generation [13].

An intelligent HRS using Restricted Boltzmann Machine (RBM)-Convolutional Neural Network (CNN) is presented in [14], which uses the big data in recommendation generation. A hybrid diagnosis scheme for coronary artery disease is presented in [15], with neural network. Similarly, a Fuzzy analytic hierarchy process (Fuzzy\_AHP) technique is presented towards cardiac disease which estimates global weights for various features of individual and classification is performed with ANN [16]. A Congenital Heart Disease diagnosis scheme is presented in [17], which uses BPNN towards recommendation.

Towards providing security for software components a Mamdani fuzzy inference system is presented in [18], which evaluates different security measures and produces recommendations. A multiple kernel learning adaptive neuro fuzzy inference (MKL with ANFIS) is presented towards diagnosis of heart disease which support the classification of individuals [19]. A detailed review on drug recommendation is presented in [20], which used several articles. A relational connection based heterogeneous model is presented in [21], which uses drug, proteins and drugs with side effects. The DTIPred is discussed according to random walk and convolutional neural network. The model integrates various interactions and representations.

A cognitive intervention-based approach is designed in [22], which clubs the result of three different approaches in drug recommendations. A meta-analysis of randomized controlled trial based anti-inflammatory drugs towards cardiac disease is presented in [23]. At last, a cross sectional long term follow up scheme with transanal endorectal pull-through (TERPT) is presented in [24] to support diagnosis and surgery.

#### A. Multi Feature Drug Curing Rate Based Recommendation System with GA

The proposed multi feature drug curing rate based recommendation system reads the medical data set. According to the traces available, the method first preprocesses the data set to remove the noisy records. Further, the method finds the drugs list and their compounds to extract different features. Using the features extracted, the method applies genetic algorithm to estimate the fitness value of different drug compound to select optimal and efficient drug compound towards specific disease. According to the fitness function, the method computes the value of multi feature drug curing rate based on which the drugs are ranked to support medical practitioner. The detailed approach is presented in this section.

The architecture of proposed RMDCRSR model is presented in Fig. 1, and shows various functional components of the mode. Each of the functional stages is discussed in detail in this section.

#### B. Preprocessing

The medical data set given has been fetched here and the preprocessing algorithm finds the list of features available in the entire trace from the data set. According to the features identified, the traces are traversed to verify the features presence with value. If a trace without the feature and value is identified, then it has been removed from the data set. Such traces removed with noisy records are used to perform feature extraction.

Consider the data set given is Meds, which contains  $\emptyset$  records and each has k number of features, then the noise removal is performed as follows:

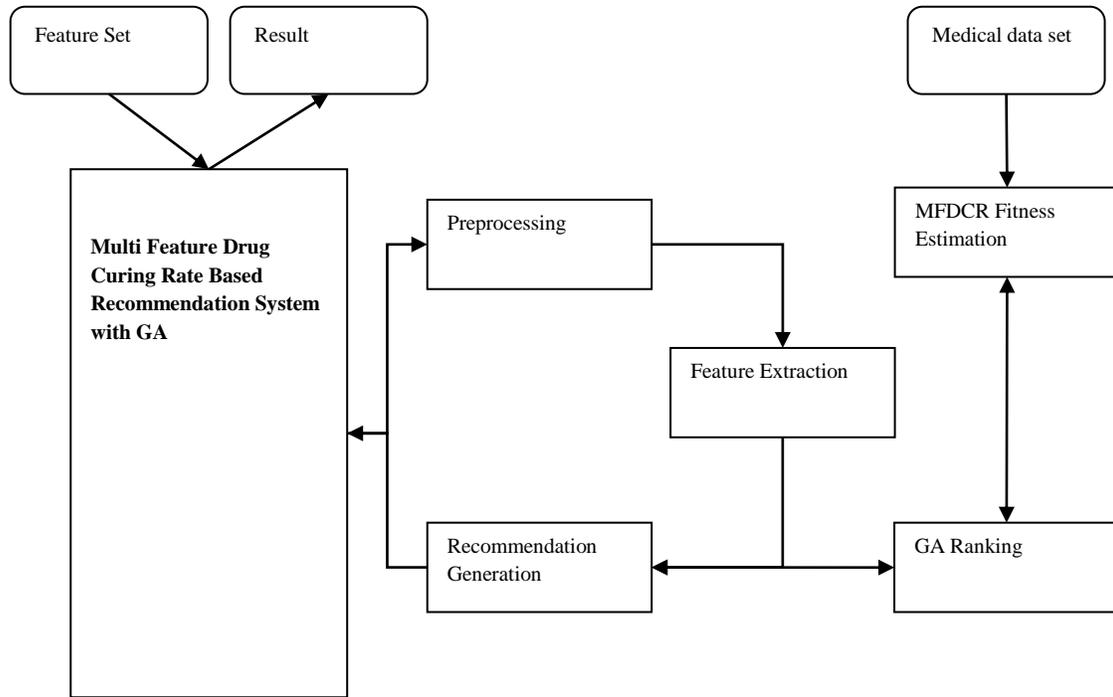


Fig. 1. Architecture of proposed MDCRSR model.

First the list of features are identified using equation (1). Features list  $F_{e\ list} =$

$$F_{e\ list} = \bigcup_{i=1}^{size(Meds)} (Meds(i).features \setminus F_{e\ list}) \cup F_{e\ list} \quad (1)$$

Now, the noisy records are identified and eliminated as follows:

$$Medsp = \bigcup_{i=1}^{size(Meds)} (if\ Meds(i) \in (\forall\ features\ (F_{e\ list}))? Medsp \cup Meds(i): Meds \cap Meds(i)) \quad (2)$$

Such noise removed data set has been used to extract the features and to support decision making.

### C. Feature Extraction

The data set preprocessed contains number of features and this phase extract several features from the data set. First, the traces belong to specific disease are identified and separated. With the traces separated, the set of drugs given for the disease has been identified and their subsequent compounds are identified. For example, for arrhythmic disease number of drugs would be used and at each drug there would be different manufacturer and volume of drug available, the feature extraction module, finds such drugs and volumes in distinct manner and collects the drugs and compounds. Identified drugs list and compounds are used to perform drug analysis to support recommendation generation. Medical Data set is considered as  $P_{mds}$  and drug vector as  $DV$ . The proposed

system will read the medical dataset and initialize the drug list  $DI$ .

For each record  $r$  Identify set of drugs  $D_s =$

$$D_s = \bigcup_{i=1}^{size(P_{mds})} Drugs \in p_{mds}(i) \quad (3)$$

For each drug  $d$

$$DI = \bigcup_{i=1}^{size(D_s)} D_s(i) \in DI? D_s \cap D_s(i): DI \cup D_s(i) \quad (4)$$

Drug vector  $Dv = \{DI\}$

The feature extraction algorithm identifies the list of drugs according to varying compounds and volumes. According to the drugs identified, the method performs recommendation generation.

### D. MFDCR Fitness Estimation

The fitness function of the proposed model estimates the fitness value by computing the multi factor drug curing rate (MFDCR). Each drug has specific curing rate on specific disease and according to that the method computes the DCR value for each disease. Based on the value of DCR, the method computes the value of MFDCR. It has been measured according to the difference on different logs status. To measure the value of fitness, the method split the traces of disease under success and failure classes. Once the traces are split, then the fitness value is measured as follows:

Consider the population given is  $p = \{D1, D4, D7, D9, D11\}$  where  $Dx$  represent the drug name and according to that

the fitness value is measured in terms of multi feature drug curing rate (MFDCR). Similarly, consider the trace set Sdt contains the traces of drugs with success treatment and Fdt represent the traces of drugs with failure treatment, then the method computes the Drug Curing Rate as follows:

First the drug trace has been identified as follows:

$$\text{Success Drug Trace ST} = \left( \sum_{i=1}^{\text{size}(Sdt)} Sdt(i) \in Dk \right) \&\& Sdt(i).state == Success \quad (5)$$

$$\text{Failure Drug Trace FT} = \left( \sum_{i=1}^{\text{size}(Sdt)} Sdt(i) \in Dk \right) \&\& Sdt(i).state == Failure \quad (6)$$

According to the values of equation (5, 6), the value of drug curing rate is measured as follows:

$$DCR = \frac{\text{Dist}(\text{Size}(ST) - \text{Size}(FT))}{\text{size}(ST) + \text{size}(FT)} \quad (7)$$

According to the drug curing rate of different drugs, the value of fitness function as MFDCR has been measured as follows:

$$MFDCR = \frac{\sum_{i=1}^{\text{size}(DI)} DI(i).DCR}{\text{size}(DI)} \quad (8)$$

Where, DI-represent the drug list and size(DI) represent the total number of drugs. According to the value of MFDCR, the fitness of the drugs pattern or population has been verified.

#### E. GA-Ranking

The proposed modified genetic algorithm finds the drugs list and their compounds. With the set of drugs, the method initially generates the population according to the size of drug list. For the populated drug compound, the method computes the MFDCR fitness value. If the fitness value greater than the threshold, then the population is selected as the result as recommendation. Otherwise, the process is iterated by generating different combinations by crossover operation and mutation operations. In this algorithm, the crossover and mutation operations are generated by computing the combinations with the number of drugs available in the earlier population. At each population, the method computes the value of MFDCR value. Finally, the method is iterated till an optimal drug composition is selected according to the value of MFDCR and threshold. Also, the method tries to reduce the number of compositions or drugs by choosing the drug population according to the MFDCR value. Finally, top few drug compositions or populations are populated as result to the user.

The proposed method Takes the Medical data set Meds, Disease D and returns Recommendations Rc.

When MFDCR < Th then

$$\text{Population set Ps} = \underset{i=1}{\text{GenerateCombinations}(i, p)}^{\text{size}(p)} \quad (9)$$

For each population pi

$$MFDCR = \text{Fitness}(pi)$$

If MFDCR > Th then

Population list PI = Rank Populations by MFDCR.

Recommendations Rc = PI

The genetic algorithm based ranking approach reads the data set and generates the initial population and computes the fitness value. According to the threshold the recommendation are generated and if the value of MFDCR is less than threshold, the process is iterated by performing crossover and mutation functions on the population of drug given. According to the value of MFDCR, the combinations of drug are ranked to produce recommendations to the user.

### III. RESULTS AND DISCUSSION

The proposed real time multi feature DCR based drug recommendation scheme with Genetic algorithm has been implemented and evaluated for its performance under various parameters. The evaluation has been performed according to the data set maintained by different medical organizations and the efficiency of the methods are measured on various factors. The performance of the proposed algorithm has been evaluated with MIMIC-III data set. MIMIC-III is an open source database provided by American critical care unit located at Beth Israel Deaconess Medical Center, Boston in the period of 2001 and 2012, which has been obtained from MIMIC-III. It covers variety of information from general patient details, intensive care unit features obtained, treatment details, diagnosis reports from lab, clinical and various other details at different stage with the improvement with the treatment and Data includes vital signs, medications, laboratory measurements, observations and notes charted by care providers, fluid balance, procedure codes, diagnostic codes, imaging reports, hospital length of stay, survival data, and more. In total it covers thousands of features in 26 tables which can be used to perform analysis on heart diseases.

TABLE I. ANALYSIS ON RECOMMENDATION GENERATION

Recommendation Generation Performance			
	50 Drugs	100 Drugs	200 Drugs
SWRL	73	79	83
HyperRecSysPA	77	82	85
MES-GRS-GA	81	86	89
MFDCR_GA	87	92	97

#### A. Performance

In the Table I performance of recommendation generation has been measured at different number of drugs cases. In each test case, the proposed MFDCR\_GA approach has produced higher recommendation performance than other approaches.

The performance of different methods in recommendation generation has been measured and presented in Fig. 2. The proposed MFDCR\_GA has produced higher performance at different number of drugs cases.

In Table II false ratio of recommendation generation has been measured for different approaches and the proposed MFDCR\_GA based approach has produced less false ratio than other approaches.

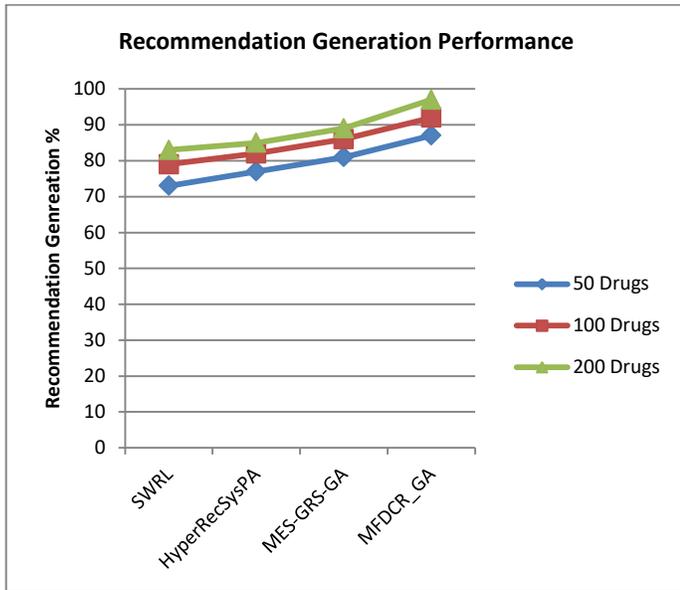


Fig. 2. Performance on recommendation generation.

TABLE II. ANALYSIS ON FALSE RATIO

False Ratio in Recommendation Generation			
	50 Drugs	100 Drugs	200 Drugs
SWRL	27	21	17
HyperRecSysPA	23	18	15
MES-GRS-GA	19	14	11
MFDCR_GA	13	8	3

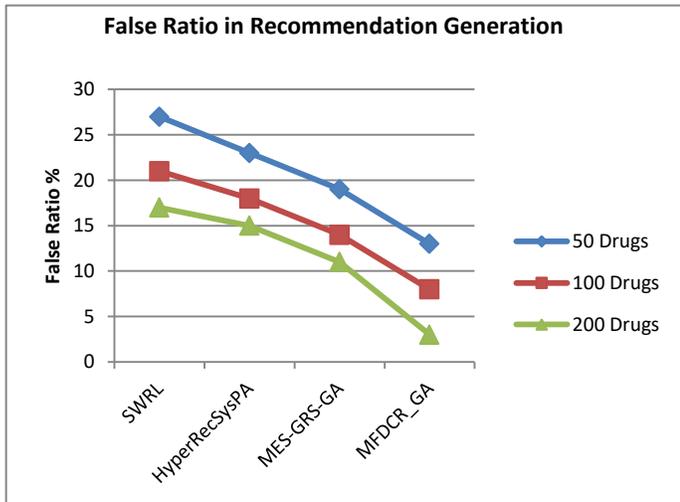


Fig. 3. False ratio in recommendation generation.

The ratio of false recommendation produced by the methods are measured and presented in Figure 3. In each test case, the proposed MFDCR\_GA approach has produced less false ratio than other approaches.

In the table 3 time complexity in recommendation generation has been measured for different approaches and the proposed MFDCR\_GA based approach has produced less time complexity than other approaches.

TABLE III. TIME COMPLEXITY

Time Complexity in Recommendation Generation			
	50 Drugs	100 Drugs	200 Drugs
SWRL	57	71	96
HyperRecSysPA	53	68	87
MES-GRS-GA	46	62	81
MFDCR_GA	23	28	39

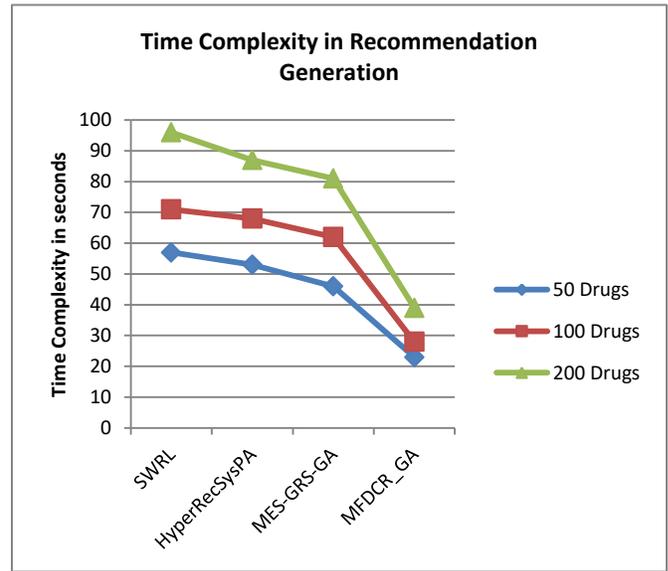


Fig. 4. Analysis on time complexity.

The time complexity in recommendation generation has been measured and presented in Fig. 4, where the proposed MFDCR\_GA has produced less time complexity than other approaches.

#### IV. CONCLUSION

This paper presented a detailed implementation of multi feature DCR recommendation system with Genetic algorithm. The method reads the data set and finds set of drugs for different disease class. Further the value of drug curing rate is measured on each drug towards various disease classes. Also, the method computes the multi feature drug curing rate (MFDCR) towards various disease classes. According to the value of MFDCR the drugs and their compounds are ranked to generate recommendation to the medical practitioners. The proposed approach has been evaluated with MIMIC-III database and the method improves the performance of recommendation generation with least false ratio and time complexity.

#### REFERENCES

- [1] Md. Deloar Hossain; Md. Shafiu Azam; Md Jahan Ali; Hakilo Sabit, Drugs Rating Generation and Recommendation from Sentiment Analysis of Drug Reviews using Machine Learning, IEEE, Emerging Technology in Computing, Communication and Electronics (ETCCE), 2020.
- [2] C. Chen, L. Zhang, X. Fan, Y. Wang, C. Xu and R. Liu, "A epilepsy drug recommendation system by implicit feedback and crossing recommendation", Proceedings – 2018 IEEE Smart World Ubiquitous

- Intelligence and Computing Advanced and Trusted Computing Scalable Computing and Communications Cloud and Big Data Computing Internet of People and Smart City Innovations Smart World IUICIATCIScaIComICBDCo, 2018.
- [3] D. Chen, D. Jin, T. T. Goh, N. Li and L. Wei, "Context-Awareness Based Personalized Recommendation of Anti-Hypertension Drugs", *J. Med. Syst.*, vol. 40, no. 9, pp. 1-10, Sep. 2016.
- [4] AThi Ngoc Trang Tran, *Recommender systems in the healthcare domain: state-of-the-art and research issues*, Springer, Journal of Ambient Intelligent Information Systems, 2020.
- [5] Atas, M., Tran, T.N.T., Felfernig, A., Polat-Erdeniz, S., Samer, R., & Stettinger, M. (2019). Towards similarity-aware constraint-based recommendation. In *Advances and trends in artificial intelligence, lecture notes in computer science*, (pp. 287–299). Springer.
- [6] Bankhele, S., Mhaske, A., & Bhat, S. (2017). V., s.: a diabetic healthcare recommendation system. *International Journal of Computer Applications*, 167, 14–18.
- [7] Castro, J., Quesada, F.J., Palomares, I., & Martínez-López, L. (2015). A consensus-driven group recommender system. *International Journal of Intelligent Systems*, 30(8), 887–906.
- [8] Colberg, S., Sigal, R., Yardley, J., Riddell, M., Dunstan, D., Dempsey, P., Horton, E., Castorino, K., & Tate, D. (2016). Physical activity/exercise and diabetes: a position statement of the american diabetes association. *Diabetes Care*, 39, 2065–2079.
- [9] Gujar, D., Biyani, R., Bramhane, T., Bhosale, S., & Vaidya, T.P. (2018). Disease prediction and doctor recommendation system. *International Research Journal of Engineering and Technology (IRJET)*, 5, 3207–3209.
- [10] Luciano Rodrigo Ferretto; Ericles Andrei Bellei, *A Physical Activity Recommender System for Patients With Arterial Hypertension*, *IEEE Access*, Vol.8, 2020.
- [11] Ritu Meena and Sonajharia Minz, *Group Recommender Systems – An Evolutionary Approach Based on Multi-expert System for Consensus*, De Gruyter, 2018.
- [12] Yar Muhammad, Muhammad Tahir, *Early and accurate detection and diagnosis of heart disease using intelligent computational model*, *scientific reports*, 2020.
- [13] Subhashini narayanan, *A novel recommender system based on FFT with machine learning for predicting and identifying heart diseases*, *Neural Computing and Applications*, 2019.
- [14] Abhaya Kumar Sahoo, *DeepReco: Deep Learning Based Health Recommender System Using Collaborative Filtering*, *MDPI, Computation*, 7(2), 2019.
- [15] Arabasadi, Z., Alizadehsani, R., Roshanzamir, M., Moosaei, H. & Yarifard, A. A. Computer aided decision making for heart disease detection using hybrid neural network-Genetic algorithm. *Comput. Methods Programs Biomed.* 141, 19–26 (2017).
- [16] Samuel, O. W., Asogbon, G. M., Sangaiah, A. K., Fang, P. & Li, G. An integrated decision support system based on ANN and Fuzzy\_AHP for heart failure risk prediction. *Expert Syst. Appl.* 68, 163–172 (2017).
- [17] Vanisree, K. & Singaraju, J. Decision support system for congenital heart disease diagnosis based on signs and symptoms using neural networks. *Int. J. Comput. Appl.* 19, 6–12 (2015).
- [18] Nazir, S., Shahzad, S., Mahfooz, S. & Nazir, M. Fuzzy logic based decision support system for component security evaluation. *Int. Arab J. Inf. Technol.* 15, 224–231 (2018).
- [19] Manogaran, G., Varatharajan, R. & Priyan, M. Hybrid recommendation system for heart disease diagnosis based on multiple kernel learning with adaptive neuro-fuzzy inference system. *Multimedia Tools Appl.* 77, 4379–4399 (2018).
- [20] Idris Rabi, *Recommender System Based on Temporal Models: A Systematic Review*, *MDPI, Applied Sciences*, 10(7), 2020.
- [21] Xiaoqiang Xu Ping Xuan, *Infering drug target interaction based randomwalk and convolution neural network*, *IEEE/ACM Trans Comput Biol Bioinform* 2021.
- [22] Zana Bućinca, Maja Barbara Malaya, and Krzysztof Z. Gajos. *To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making*. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1), 2021.
- [23] Ivan Wudexi, Elica Shokri, *Comparative Effectiveness of Anti-Inflammatory Drug Treatments in Coronary Heart Disease Patients: A Systematic Review and Network Meta-Analysis*, *Hindawi, Mediators of Inflammation*, 2021.
- [24] Johan Hasseri, Josefine Hedbys, *Treatment and Patient Reported Outcome in Children with Hirschsprung Disease and Concomitant Congenital Heart Disease*, *Hindawi, Biomedical research international*, 2017.

# The Predictors of Mobile Banking Usage: A Systematic Literature Review

Mohammed Abd Al-Munaf Hashim, Prof. Dr. Zainuddin bin Hassan

College of Computing and Informatics (CCI), Universiti Tenaga Nasional, Kuala Lumpur, Malaysia

**Abstract**—Mobile banking has become an essential method to conduct banking transaction. However, number of users worldwide are still limited. The purpose of this study is to review the literature and understand the status of m-banking adoption, usage, and loyalty. Keywords were used to search for related articles in three databases namely, Web of Science (WoS), Scopus, and Google scholar. Filtering process was conducted to select the most related articles. This has resulted in reviewing 45 articles. The findings showed that number of articles pertaining to m-banking is increasing. Malaysia and Indonesia have the largest number of articles. The technology acceptance model (TAM) is being used widely in the m-banking literature and most of the reviewed studies are empirical with adequate sample size. This explains the increased usage of structural equation model (SEM). The most critical factors for m-banking adoption, usage, and loyalty are service quality, trust, perceived usefulness, perceived ease of use, security, risk, privacy, and social influence. Future research is suggested to examine the m-banking in different region and using mediating and moderating variables to explain the variation in the adoption.

**Keywords**—M-banking; TAM; Service quality; Loyalty; UTAUT

## I. INTRODUCTION

Mobile banking is a new technology that enables the access and the execution of financial transaction using mobile device. This trend has been increasingly used by users around the world with varied percentage.

In developed countries, it was noted that three out of ten are using m-banking to settle their financial transaction. The percentage reduced in other non-developed countries. M-banking is beneficial for users and the banks as it reduces the physical effort of users and enable them to conduct their transactions without the need to visit banks or going through routines and paperwork. On the other hand, the m-banking speed the service process and reduce the workload of employees at banks and it also reduces the financial operational cost for banks. However, it has been observed that the use of m-banking by users is still limited and more studies are needed to understand the application and the adoption of users regarding the use of m-banking.

Previous studies with the opinion that the perceived usefulness (PU) and perceived ease of use (PEOU) are critical factors for the adoption [1, 2, 3]. On the other hand, the privacy, security, and trust were identified as the most important variables in the context of m-banking [4, 5, 6]. Other studies related the usage to the service quality [7, 8]. Literature regarding the critical factors that lead to the usage of m-

banking has no agreement on the variables that can be deployed to enhance the utilization of m-banking. However, it has been observed that the use of m-banking by users is still limited and more studies are needed to understand the application and the adoption of users regarding the use of m-banking. Theoretical framework has been deployed by previous studies to enhance the explanatory power of the m-banking. One of the widely used theoretical models is the technology acceptance model (TAM) by Davis [9].

TAM is with the opinion that the adoption is mainly related to the perception of users regarding the PEOU and the PU of the technology. On the same context, one of the emerging models is the unified theory of acceptance and use of technology (UTAUT) by Venkatesh et al. [10]. UTAUT deployed four main variables and considered these variables as critical for the adoption of any new technology. These variables include the effort expectancy (EE), performance expectancy (PE), social influence (SI), and facilitating condition (FC). Other theoretical models are used also in the literature. These include the information system success (IS) success which relates the usage of a new technology to the quality of service, system, and information as well as the satisfaction of users.

There is no agreement in the literature regarding the predictors of m-banking or the theoretical model that explain the loyalty to use m-banking. Previous studies are in general empirical and there is a need for a review to understand the status of the literature. Accordingly, this study aims to review the literature and identify the most critical theories and predictors in the context of m-banking. The study also aims to provide the stakeholders with the research gaps and the directions of future works. Based on the fact that this study is a literature review study, the next section discusses the methodology of conducting the systematic literature review (SLR) followed by a summary of the reviewed studies. The findings are discussed as well as the limitation and direction of future work.

## II. RESEARCH METHODOLOGY

This study is an SLR that aims to understand the status of the literature and identify the critical factors for adopting m-banking. To fulfil this objective, the study used specific keywords to look for articles. The keywords include m-banking, predictor of m-banking, m-banking usage, and theories of m-banking. Reliable databases were searched to extract the articles (Fig. 1). Web of Science (WoS), Scopus, and google scholar were used to find the articles.

A total of 399 articles were found related to the issue. However, to have an updated view, the articles between 2016 and 2021 were selected. This has reduced the articles to 187. Further filtering using the language i.e., English language or the scope i.e., theoretical article and remove technical articles. This has further reduced the articles to 76 articles. A full reading was conducted on these 76 articles, and this has resulted in 45 articles that are related to the m-banking adoption.

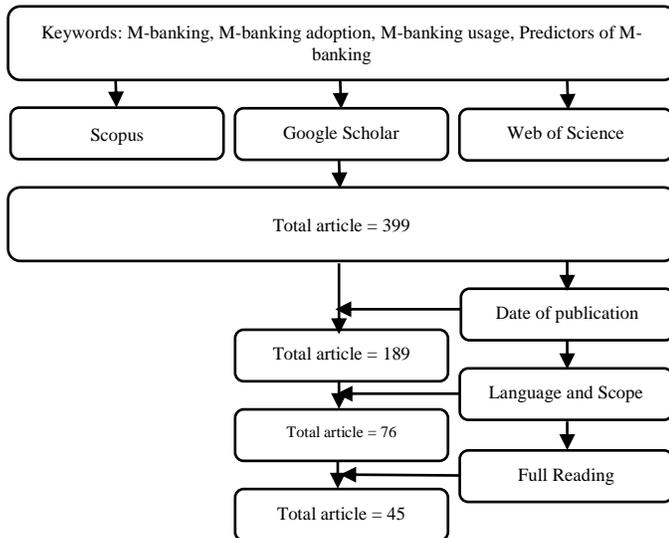


Fig. 1. Process of selecting the articles.

### III. SUMMARY OF REVIEWED STUDIES

The reviewed studies can be divided into categories based on the theoretical framework. Previous studies deployed TAM. For example, Zhao, Chen and Wang deployed TAM to examine the satisfaction and loyalty of using M-banking [1]. The findings showed that social influence and the psychological ownership has greater effect than the variables of TAM. Similarly, the findings of Rubiah Abu Bakar indicated that variables such as security and privacy are more critical than the PEOU of m-banking [11]. Yuan et al. also deployed TAM to examine the effect of satisfaction and technology fit along with variables of TAM. The findings indicated that satisfaction, PU, task-technology fit, and risk positively associated with customer' loyalty [2]. Moreover, the results showed that confirmation, PEOU, and task-technology fit positively impact PU.

Yun Min Low deployed TAM to examine the loyalty of using m-banking. The findings showed that subjective norms, PEOU and convenience risk are critical factors for the usage of loyalty [12]. Liébana-Cabanillas et al. found that customer loyalty has been significantly influenced by satisfaction [13]. Munoz-Leiva et al. and found that TAM can explain the intention to use the M-banking [14]. Abdinoor and Mbamba found that the TAM model integrated with cost and awareness are useful in explaining the intention to use M-banking [15]. Aliza Kasim found that variables of TAM and facilitating conditions are positively influencing the customer intention to use M-banking [16]. Priya, Gandhi et al. utilized TAM and found that customer's behavioural intention and satisfaction in

using the service are significantly influencing by PU, PEOU, perceived credibility, and structural assurance [17]. Moreover, the positive mediating effect of customer satisfaction has confirmed on the relationship among the related factors and behavioural intention. Meanwhile, the effect of perceived risk on behavioural intention to use the service has been reported to have insignificant effect.

Mostafa and Eneizan indicated that facilitating conditions and self- efficacy are positively affecting on PEOU and PU factors. Furthermore, the results have indicated a positive impact of PEOU and PU on costumer's behavioural intentions to accept and use M-banking technology [18]. Nawaz and Yamin utilized TAM and found that PEOU and PU have a direct positive impact on the behavioral intention to use the service. Trust found to have a positive impact on customer behavioral intention to use m-banking [19]. Mutahar proposed perceived value as a moderator in among the variables of TAM [20]. Ezeh and Nwankwo utilized TAM and found that PEOU, financial cost and perceived credibility have significant effect on intention to use m-banking [21]. Malaquias and Hwang also found that PEOU and PU as well as trust and social influence affect the actual use of m-banking [22]. Mulia et al. added the customer intimacy as a variable and deployed TAM to examine the user satisfaction and loyalty of using m-banking. The findings showed that Customer intimacy affected the directly the customer satisfaction and the loyalty [23].

Another theoretical framework that has been used by the reviewed studies is the unified theory of acceptance and use of technology (UTAUT). De Sena Abrahão et al. deployed UTATU and included variables such as cost and risk to examine the behavioural intention to use m-banking [24]. The findings showed that large portion of the variation can be explained by UTAUT. Maduku found that variables of UTAUT as well as self-efficacy and structural awareness affected positively the behavioural intention to use m-banking [25]. Zendeudel, Paim et al. found that PE, EE, cost and trust are significantly affecting the customer intention to use M-banking [26]. Singh and Srivastava found that trust, security, privacy, and social influence affected the intention to use M-banking [27]. Baabdullah, Alalwan et al. found that UTAUT variables can explain the loyalty to use M-banking applications [28].

The third theoretical model that has been used is the IS success. This model was deployed to emphasize on the service, system, and information quality. For instance, Mohd Thas Thinker, Amin et al. deployed the model to explain the loyalty of using m-banking. The findings showed that usability, customer service, satisfaction and trust in m-banking service influencing the continuance intention or loyalty toward using m-banking service. In addition, continuance usage or loyalty is positively affected by trust and customer satisfaction mediator factors [29]. Sharma and Sharma uses the model also to explain the actual use of m-banking. The study found that satisfaction and intention to use are significantly influenced by the service quality and trust which leads to influencing the actual usage of the service [30]. Expressly, a more trustable service with better quality will help in keep the current customer using the service and attract a new wave of users to join the system. Service quality theory has been used also in the literature of M-

banking. For example, Suariedewi and Suprapti deployed the theory to explain the e-loyalty of m-banking [8]. Similarly, Zhou et al. used the theory to explain the m-banking service quality and loyalty [7].

Mixed theories were used in few numbers of previous studies. Kumar, Israel et al. combined ECT and SDT to explain the continuous of usage and loyalty to m-banking. The findings showed that the continuance intention to use m-banking is influencing by the following factors: satisfaction, intrinsic regulation and identified regulations, while influenced by expectation-confirmation, trust, and quality factors effect stingingly in the customer satisfaction [31]. UTAUT and IS success theories were combined in the study of Windasari and Albashrawi and the findings based on gender suggested that service quality affected PE and EE of male but not female. Information quality affected the facilitating condition of male but not female. PE, EE, and FC affected the loyalty. Service quality affected satisfaction of female and satisfaction affected the loyalty [32]. Symbolic value theory and brand equity theory were combined in the study of Esmaeili et al., the findings showed that usability affected trust and satisfaction. Perceived risk affected loyalty and relative advantage affected customer loyalty [33].

Large number of the reviewed studies did not deploy any theory, and this can be seen in the different studies [34, 35, 36, 37, 38, 4, 39, 40, 5, 6, 41, 42, 43]. Most of these studies focus on the security, privacy, and risk of using m-banking. Table I shows a summary of the reviewed articles which include the author, years, country, sample, and theory.

TABLE I. SUMMARY OF REVIEWED STUDIES

Author/ Year	Country	Sample	Subhead
(Zhao, Chen, & Wang, 2016) [1]	Taiwan	791 users	TAM
(Susanto et al., 2016) [34]	South Korea	201	ECT
(Yuan et al., 2016) [2]	China	434	TAM
(de Sena Abrahão et al., 2016) [24]	Brazil	605	UTAUT
(Rubiah Abu Bakar 2017) [11]	Malaysia	150	TAM
(Shuhidan, Hamidi et al. 2017) [35]	Malaysia	384	-
(Yun Min Low 2017) [12]	Malaysia	261	TAM
(Liébana-Cabanillas et al. 2017) [13]	Chile	218	TAM
(Munoz-Leiva, Climent-Climent et al. 2017) [14]	Spain	218	TAM
(Chiu, Bool et al. 2017) [36]	Philippines	314	-
(Abdinoor and Mbamba 2017) [15]	Tanzania	200	TAM
(Aliza Kasim 2017) [16]	Malaysia	359	TAM
(Maduku 2017) [25]	South Africa	401	UTAUT
(Alalwan et al., 2017) [44]	Jordan	343	UTAUT2
(Kumar, Israel et al. 2018) [31]	India	744	ECT, SDT
(Masrek, Halim et al. 2018) [37]	Malaysia	365	-

Author/ Year	Country	Sample	Subhead
(Jamshidi, Keshavarz et al. 2018) [38]	Iran	927	-
(Zendehtdel, Paim et al. 2018) [26]	Malaysia	400	UTAUT
(Singh and Srivastava 2018) [27]	India	875	UTAUT2
(Priya, Gandhi et al. 2018) [17]	India	269	TAM
(Khasawneh, Hujran et al. 2018) [4]	Jordan	404	-
(Mohd Thas Thaker, Amin et al. 2018) [29]	Malaysia	250	D&M IS
(Mostafa and Eneizan 2018) [18]	Libya	261	TAM
(Wichittakul and Prasongsukam 2018) [39]	Thailand	336	-
(Nawaz and Yamin 2018) [19]	Sri Lankan	695	TAM
(Ahmed M. Mutahar 2018) [45]	Yemen	Review	TAM
(Johannes, Indarini et al. 2018) [40]	Indonesia	200	-
(Ezeh and Nwankwo 2018) [21]	Nigeria	200	TAM
(Khan, Lima et al. 2018) [5]	Bangladesh	240	-
(Baabdullah, Alalwan et al. 2019) [28]	Saudi Arabia	429	UTAUT2
(Sharma and Sharma 2019) [30]	Oman	227	D&M IS
(Malaquias and Hwang 2019) [22]	Brazil and USA	375	TAM
(Windasari & Albashrawi, 2020) [32]	US	516	UTAUT and IS success
(Purwanto et al., 2020) [6]	Indonesia	395	-
(Mulia et al., 2020) [23]	Global	300	TAM
(Tumewah et al., 2020) [41]	Indonesia	505	NIL
(Suariedewi & Suprapti, 2020) [8]	Indonesia	120	Service quality
(Khan et al., 2021) [42]	Bangladesh	362	Nil
(Esmaeili et al., 2021) [33]	Iran	411	Symbolic value theory Brand equity theory
(Zhou et al., 2021) [7]	China	224	Service quality
(Parera & Susanti, 2021) [43]	Indonesia	105	Nil

#### IV. FINDINGS

The findings of this study are derived using frequency analysis. The analysis was conducted after extracting the information of the articles. The analysis is conducted using excel sheet and it includes the year of publication, country, theoretical framework, approach, sample size, and data analysis technique. More importantly, the analysis included the most critical factors for m-banking.

##### A. Year of Publications

The year of publication is given in Fig. 2. It shows that the number of articles has increased between 2016 and 2018 and

this could be due to the notion that the application of m-banking has increased during this period. Number of articles reduced between 2019 and 2022 and this could be due to the searching criteria for articles in this period (see Fig. 2).

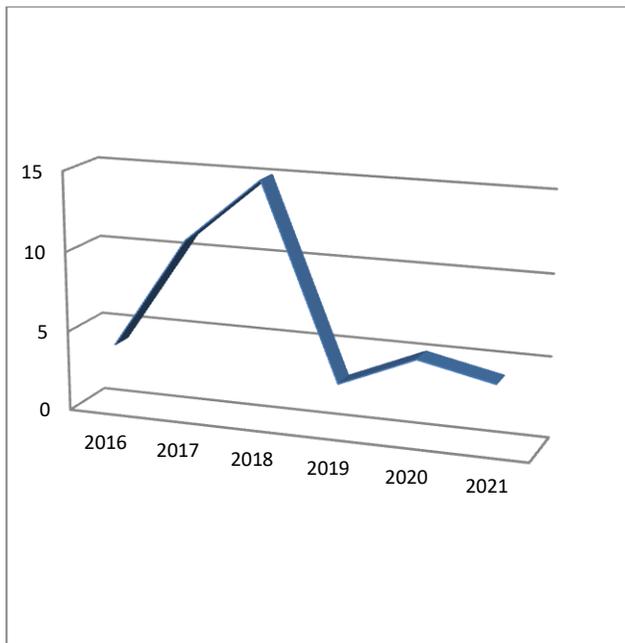


Fig. 2. Year of publications.

### B. Country of Origin

The country of origin is where the research has been conducted. It can be seen that the number of articles in Malaysia is the highest followed by Indonesia, India, and Bangladesh. This could be also due to the wide spread of M-banking by banks in the Asian region. Fig. 3 shows the distribution of articles based on country of origin.

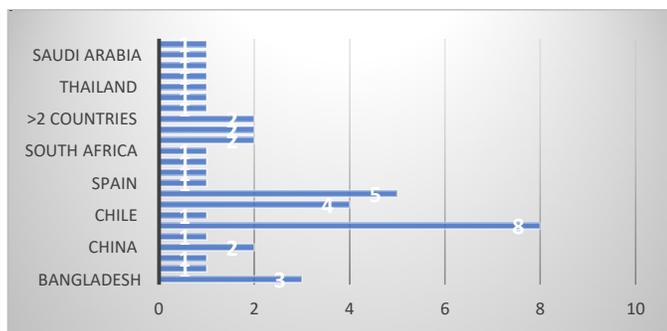


Fig. 3. Country of origin.

### C. Sample Size

The sample size of the reviewed studies is shown in Fig. 4. The minimum sample size has been used in the reviewed studies account to 105 and the maximum is 927 with mean score of 373. This means that on average, there are 373 respondents in the reviewed studies. This average indicates that there is a shift in the analytical approach. Fig. 5 shows the sample size.

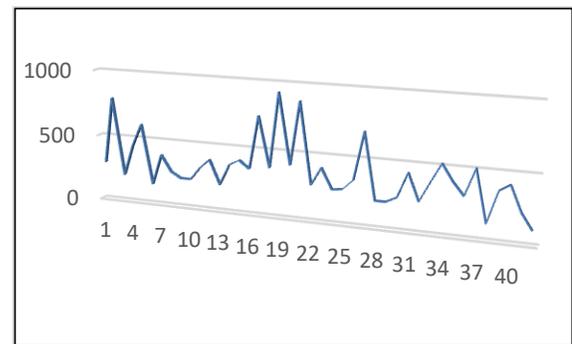


Fig. 4. Sample size.

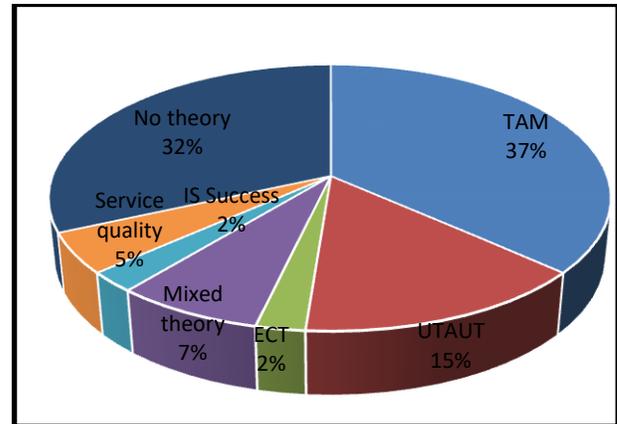


Fig. 5. Theoretical framework.

### D. Study Approach

The study approach is shown in Fig. 6. It can be seen that there are 98% of empirical studies while review studies accounted to 2%. This shows that the literature is empirical in nature and limited numbers of review studies were conducted.

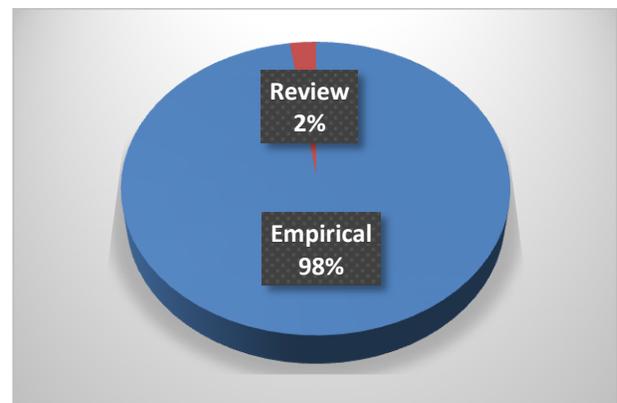


Fig. 6. Theoretical framework.

### E. Statistical Tools

The statistical tools that have been used by the reviewed studies to analyses and explain the findings are shown in Fig. 7. It can be seen the SPSS has been used by 41% of the reviewed studies followed by PLS with 32%, AMOS with 20% and LISREL with 7%. The first generation of data analysis represented by SPSS accounted to 41% while the second

generation known as structural equation modelling (SEM) and includes LISREL, AMOS, and PLS accounted to 59%. This confirms that there is a shift toward using advance analytical tools to examine the association among the variables. Again, this is also confirmed from the mean of sample size which is 373. This is because the use of SEM requires a minimum of sample size of 200.

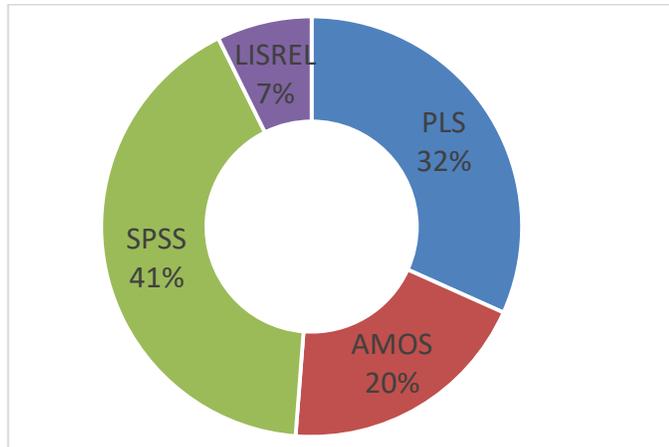


Fig. 7. Analytical techniques.

#### F. Critical Factors

Factors of the 45 articles were extracted and a frequency analysis was conducted to identify the most related articles. The findings showed that service quality is one of the most critical factors for user satisfaction and loyalty to use m-banking. This followed by trust, PU, PEOU, security, and perceived risk as well as privacy, and social influence. These variables have been identified by prior literature to be critical for the behavioral and actual behavior toward using the m-banking by customers in various countries as shown in Fig. 8.

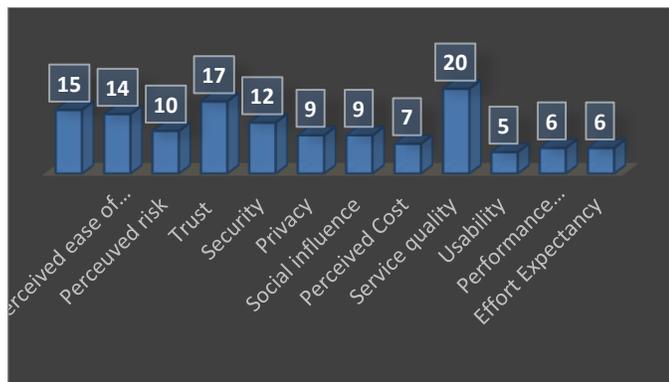


Fig. 8. Critical factors for m-banking.

#### V. DISCUSSION AND IMPLICATIONS

This study was conducted to review the literature pertaining to the m-banking usage and adoption by users. The study also aimed to identify the gaps and the direction of future work as well as the predictors of using the m-banking. The finding showed that the number of articles published is increasing and this is a signal that the awareness about m-banking is also increasing. The country of origin shows that Malaysia is a leading country in m-banking studies followed by Indonesia.

This could be due to the lifestyle of the people in these countries as online banking has become the norm and the preferred method of settling the financial transaction. In term of the theoretical model, the literature showed that TAM is still dominating the research of m-banking followed by to a less degree with UTAUT. Emerging theories such as IS success is used by limited number of studies. The quantitative approach is used widely by the literature and sample size is adequate for the analyses of the studies. In addition, the use of SEM is increasing in the literature.

In terms of the predictors, the service quality is the most important predictor of m-banking loyalty, adoption, and usage. This is followed by trust and the variable of TAM such as PEOU and PU. Security, perceived risk and privacy as well as social influence are also critical for the usage of m-banking. This finding is in line with the findings of previous literature who found that the service quality and trust as well as security and privacy along with the variables of TAM are critical for the m-banking usage [1, 2, 12, 8, 45].

Accordingly, this study suggested that these factors can represent the predictors of users of m-banking across various countries and region. In addition, the study suggests that TAM is one of the most used theoretical frameworks. Therefore, decision makers are suggested to focus on the service provided by the m-banking application. Creating trust between the banks and the customers is essential for enhancing the intention to use the application. In addition, the security procedures is vital for the increase adoption of m-banking. The PU and the ease of using the application can be important variables for the adoption of m-banking.

#### VI. CONCLUSION, LIMITATION, AND FUTURE WORK

This study has been conducted to describe the literature about the m-banking adoption and to understand the predictors and theoretical foundation of the literature. The findings showed that there is an increase in the awareness and academic research into m-banking. The finding also showed that countries such as Malaysia, Indonesia, and India have received largest number of research into m-banking. TAM is still the dominating theoretical framework followed by UTAUT with empirical approach is overwhelming in the literature. The sample size is sufficient and larger than 200 responses which indicate that there is a shift toward using SEM which requires more than 200 responses. More importantly, the findings showed that service quality is the most important and widely frequent variable in the literature followed by trust, PU, PEOU, security, risk, social influence, and privacy.

The study reviewed 45 articles extracted from Scopus, Woos, and Google scholar. The number of articles is limited due to the searching criteria, scope, and year of publication. Future studies can expand the scope, year of publication to increase the number of the reviewed article and the generalization of the findings. The findings showed that there is limited number of review studies and therefore, it is suggested for more studies to review the literature of m-banking. In term of the year of publication, number of articles between 2019 and 2021 has decreased and researchers are suggested to conduct more studies and especially in the time of COVID-19 where the online banking has become essential for

all individual and organization to finalize their financial transaction. Research into m-banking is high in Asian countries such as Malaysia and Indonesia. However, few studies observed in other countries. Thus, more studies are suggested in other countries and region such as the Middle East, and Africa to have clearer view about the m-banking worldwide.

In term of theoretical framework, the TAM is widely used followed by UTAUT. However, mixed theories or combined theories have been used by a few numbers of studies. Therefore, it is suggested that the future research should include the more theories such as the TAM and social exchange theory or TAM and IS success, UTAUT and IS success and these theories can explain more the aviator in technology adoption such as m-banking. It was also noted that few of the previous studies have deployed mediator or moderator. Thus, future studies can incorporate moderating variables such as trust, innovativeness, education, gender, and IT knowledge. Mediating variables can be the enjoyment and playfulness of m-banking.

The critical variables identified in this study include service quality, trust, PU, PEOU, risk, security, social influence, and privacy. Future researcher can examine empirically the effect of these variables on the m-banking adoption to provide the decision makers with empirical view that can lead to more usage of m-banking.

#### ACKNOWLEDGMENT

I am deeply grateful to my supervisor Prof. Dr. Zainuddin bin Hassan, for providing invaluable guidance, support, and encouragement throughout my research. His expertise and knowledge have been instrumental in shaping the direction and outcome of this study. I also extend my heartfelt thanks to my family and friends who have been a constant source of love and support during this journey. Their unwavering belief in me has been a source of strength, and I am grateful to have them in my life.

This research would not have been possible without the support and encouragement of these individuals. I am eternally grateful for their contributions, and I dedicate this work to them.

#### REFERENCES

- [1] Zhou, M., Zhao, L., Kong, N., Campy, K. S., Xu, G., Zhu, G., ... & Wang, S. (2020). Understanding consumers' behavior to adopt self-service parcel services for last-mile delivery. *Journal of Retailing and Consumer Services*, 52, 101911.
- [2] Yuan, S., et al. (2016). "An investigation of users' continuance intention towards mobile banking in China." *Information Development* 32(1): 20-34.
- [3] Johannes, V. D., et al. (2018). Usability, customer satisfaction, service, and trust towards mobile banking user loyalty. 15th International Symposium on Management (INSYMA 2018), Atlantis Press.
- [4] Khasawneh, M. H. A., et al. (2018). "A quantitative examination of the factors that influence users' perceptions of trust towards using mobile banking services." *International Journal of Internet Marketing and Advertising* 12(2): 181-207.
- [5] Khan, A. G., et al. (2018). "Understanding the Service Quality and Customer Satisfaction of Mobile Banking in Bangladesh: Using a Structural Equation Model." *Global Business Review*: 0972150918795551.
- [6] Purwanto, E., Deviny, J., & Mutahar, A. M. (2020). The Mediating Role of Trust in the Relationship between Corporate Image, Security, Word of Mouth and Loyalty in M-Banking Using among the Millennial Generation in Indonesia. *Management and Marketing*, 15(2), 255–274. <https://doi.org/10.2478/mmcks-2020-0016>.
- [7] Zhou, Q., Lim, F. J., Yu, H., Xu, G., Ren, X., Liu, D., Wang, X., Mai, X., & Xu, H. (2021). A study on factors affecting service quality and loyalty intention in mobile banking. *Journal of Retailing and Consumer Services*, 60(June 2020), 102424. <https://doi.org/10.1016/j.jretconser.2020.102424>.
- [8] Suariedewi, I. G. A. A. M., & Suprpti, N. W. S. (2020). Effect of mobile service quality to e-trust to develop e-satisfaction and e-loyalty mobile banking services. *International Research Journal of Management, IT and Social Sciences*, 7(1), 185–196. <https://doi.org/10.21744/irjmis.v7n1.836>.
- [9] Davis, F. D. (1989). Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. *Source: MIS Quarterly*, 13(3), 319–340.
- [10] Venkatesh, Morris, Davis, & Davis. (2003). User Acceptance of Information Technology: Toward a Unified View. *MIS Quarterly*. <https://doi.org/10.2307/30036540>.
- [11] Rubiah Abu Bakar, N. A. A., Adida Muhammad, Mazura Muda (2017). "PERCEIVED EASE OF USE, SECURITY AND PRIVACY OF MOBILE BANKING." *International Journal of Business, Economics and Law* 13(2).
- [12] Yun Min Low, C. F. G., Amran Rasli (2017). "USERS' LOYALTY TOWARDS MOBILE BANKING IN MALAYSIA." *Journal of Internet Banking and Commerce* 22,(S7).
- [13] Liébana-Cabanillas, F., et al. (2017). "Unobserved heterogeneity and the importance of customer loyalty in mobile banking." *Technology Analysis & Strategic Management* 29(9): 1015-1032.
- [14] Munoz-Leiva, F., et al. (2017). "Determinants of intention to use the mobile banking apps: An extension of the classic TAM model." *Spanish journal of marketing-ESIC* 21(1): 25-38.
- [15] Abdinoor, A. and U. O. Mbamba (2017). "Factors influencing consumers' adoption of mobile financial services in Tanzania." *Cogent Business & Management* 4(1): 1392273.
- [16] Aliza Kasim, S. A., Rosli Mahathir (2017). "A Study of Behavioral Intention to Use Mobile Banking in Malaysia." *International Journal of Advanced Studies in Social Science & Innovation* 1(1).
- [17] Priya, R., et al. (2018). "Mobile banking adoption in an emerging economy: An empirical analysis of young Indian consumers." *Benchmarking: An International Journal* 25(2): 743-762.
- [18] Mostafa, A. A. and B. Eneizan (2018). "Factors Affecting Acceptance of Mobile Banking in Developing Countries." *International Journal of Academic Research in Business and Social Sciences* 8(1): 340-351.
- [19] Nawaz, S. S. and F. B. M. Yamin (2018). "Sri Lankan customers' behavioural intention to use mobile banking: a structural equation modelling approach."
- [20] Ahmed M. Mutahar, O. I., Abhijit Ghosh, Ahmed Hamoud Al-Shibami (2018). "Perceived value as a moderator variable in mobile banking context: an extension of technology."
- [21] Ezeh, P. C. and N. Nwankwo (2018). "Factors that Influence the Acceptance of Mobile Money in Nigeria." *Journal of Research in Marketing* 8(2): 684-697.
- [22] Malaquias, R. F. and Y. Hwang (2019). "Mobile banking use: A comparative study with Brazilian and US participants." *International Journal of Information Management* 44: 132-140.
- [23] Mulia, D., Usman, H., & Parwanto, N. B. (2020). The role of customer intimacy in increasing Islamic bank customer loyalty in using e-banking and m-banking. *Journal of Islamic Marketing*, 12(6), 1097–1123. <https://doi.org/10.1108/JIMA-09-2019-0190>.
- [24] de Sena Abrahão, R., et al. (2016). "Intention of adoption of mobile payment: An analysis in the light of the Unified Theory of Acceptance and Use of Technology (UTAUT)." *RAI Revista de Administração e Inovação* 13(3): 221-230.

- [25] Maduku, D. K. (2017). "Customer acceptance of mobile banking services: Use experience as moderator." *Social Behavior and Personality: an international journal* 45(6): 893-900.
- [26] Zendeheel, M., et al. (2018). "A STUDY OF MOBILE BANKING IN MALAYSIA BASED ON THE UTAUT MODEL." *NATIONAL ACADEMY OF MANAGERIAL STAFF OF CULTURE AND ARTS HERALD*(1): 111-118.
- [27] Singh, S. and R. Srivastava (2018). "Predicting the intention to use mobile banking in India." *International Journal of Bank Marketing* 36(2): 357-378.
- [28] Baabdullah, A. M., et al. (2019). "Consumer use of mobile banking (M-Banking) in Saudi Arabia: Towards an integrated model." *International Journal of Information Management* 44: 38-52.
- [29] Mohd Thas Thaker, M. A. B., et al. (2018). "What keeps Islamic mobile banking customers loyal?" *Journal of Islamic Marketing*.
- [30] Sharma, S. K. and M. Sharma (2019). "Examining the role of trust and quality dimensions in the actual usage of mobile banking services: An empirical investigation." *International Journal of Information Management* 44: 65-75.
- [31] Kumar, R. R., et al. (2018). "Explaining customer's continuance intention to use mobile banking apps with an integrative perspective of ECT and Self-determination theory." *Pacific Asia Journal of the Association for Information Systems* 10(2).
- [32] Windasari, N. A., & Albashrawi, M. (2020). Behavioral routes to loyalty across gender on m-banking usage. *Review of International Business and Strategy*, 31(3), 339–354. <https://doi.org/10.1108/RIBS-06-2020-0073>.
- [33] Esmaceli, A., Haghgoo, I., Davidaviciene, V., & Meidute-Kavaliauskiene, I. (2021). Customer loyalty in mobile banking: Evaluation of perceived risk, relative advantages, and usability factors. *Engineering Economics*, 32(1), 70–81. <https://doi.org/10.5755/j01.ee.32.1.25286>.
- [34] Susanto, A., et al. (2016). "Determinants of continuance intention to use the smartphone banking services: an extension to the expectation-confirmation model." *Industrial Management & Data Systems* 116(3): 508-525.
- [35] Shuhidan, S. M., et al. (2017). *Perceived Risk towards Mobile Banking: A case study of Malaysia Young Adulthood*. IOP Conference Series: Materials Science and Engineering, IOP Publishing.
- [36] Chiu, J. L., et al. (2017). "Challenges and factors influencing initial trust and behavioral intention to use mobile banking services in the Philippines." *Asia Pacific Journal of Innovation and Entrepreneurship* 11(2): 246-278.
- [37] Masrek, M. N., et al. (2018). "The Impact of Perceived Credibility and Perceived Quality on Trust and Satisfaction in Mobile Banking Context." *Asian Economic and Financial Review* 8(7): 1013-1025.
- [38] Jamshidi, D., et al. (2018). "Mobile banking behavior and flow experience: An integration of utilitarian features, hedonic features and trust." *International Journal of Social Economics* 45(1): 57-81.
- [39] Wichittakul, C. and K. Prasongsukarn (2018). Factors affecting the level of trust in mobile banking: A case study of customer perception toward commercial mobile banking adoption in Bangkok, Thailand. 2018 5th International Conference on Business and Industrial Research (ICBIR), IEEE.
- [40] Johannes, V. D., et al. (2018). Usability, customer satisfaction, service, and trust towards mobile banking user loyalty. 15th International Symposium on Management (INSYMA 2018), Atlantis Press.
- [41] Tumewah, E., Juniarta, & Kurniawan, Y. (2020). The Effect of M-Banking Service Quality and Customer Perceived Value to Satisfaction and Loyalty of Bank XYZ Customers. *International Journal of Management and Humanities*, 4(6), 132–138. <https://doi.org/10.35940/ijmh.f0634.024620>.
- [42] Khan, M. R., Rana, S., & Hosen, M. I. (2021). Impact of Trustworthiness on the Usage of M-banking Apps: A Study on Bangladeshi Consumers. *Business Perspectives and Research*. <https://doi.org/10.1177/22785337211001969>.
- [43] Parera, N. O., & Susanti, E. (2021). Customer loyalty based on mobile banking usability. 2(1), 39–48.
- [44] Alalwan, A. A., et al. (2017). "Factors influencing adoption of mobile banking by Jordanian bank customers: Extending UTAUT2 with trust." *International Journal of Information Management* 37(3): 99-110.
- [45] Mutahar, A. M., Isaac, O., Ghosh, A., & Al-Shibami, A. H. (2018). Perceived value as a moderator variable in mobile banking context: an extension of technology acceptance model (TAM). *International Journal of Management and Human Science (IJMHS)*, 2(1), 1-8.

# A Novel Approach: Tokenization Framework based on Sentence Structure in Indonesian Language

Johannes Petrus<sup>1</sup>, Ermatita<sup>2\*</sup>, Sukemi<sup>3</sup>, Erwin<sup>4</sup>

Informatics, Universitas Multi Data Palembang, Palembang, Indonesia<sup>1</sup>  
Department of Computer Science, Universitas Sriwijaya, Palembang, Indonesia<sup>2,3,4</sup>

**Abstract**—This study proposes a new approach in the sentence tokenization process. Sentence tokenization, which is known so far, is the process of breaking sentences based on spaces as separators. Space-based sentence tokenization only generates single word tokens. In sentences consisting of five words, tokenization will produce five tokens, one word each. Each word is a token. This process ignores the loss of the original meaning of the separated words. Our proposed tokenization framework can generate one-word tokens and multi-word tokens at the same time. The process is carried out by extracting the sentence structure to obtain sentence elements. Each sentence element is a token. There are five sentence elements that is Subject, Predicate, Object, Complement and Adverbs. We extract sentence structures using deep learning methods, where models are built by training the datasets that have been prepared before. The training results are quite good with an F1 score of 0.7 and it is still possible to improve. Sentence similarity is the topic for measuring the performance of one-word tokens compared to multi-word tokens. In this case the multiword token has better accuracy. This framework was created using the Indonesian language but can also use other languages with dataset adjustments.

**Keywords**—Token; tokenization; multi-word; sentence structure; sentence elements

## I. INTRODUCTION

In the current era, the amount of information is increasing very rapidly [1], a lot of information is available in text form from various types of documents such as magazines, e-books, research results, social media, emails, pdf files, video, audio, images, and large amounts of business content. Experts predict the volume of text documents will grow by 80% by 2025. To be useful, text data must be processed into information with text mining techniques [2].

To be processed, text data needs to be prepared at the text-pre-processing stage. This stage is the first important step of any data mining process to achieve better accuracy [3]. This process will change the data from its original form into a form that is easier to observe and explore [4]. One of the activities in pre-processing is tokenization besides case folding, filtering/stop-words removal, lemmatization, stemming [5], [6] including normalization and removing irrelevant words [7]. Stopwords are the least important words in a sentence, and ignoring them can help identify the most important words [8].

Tokenization is a fundamental process in almost all Natural Language Processing applications. The standard approach is single-word tokenization, in which the input string is split word

by word using spaces as separators [9]. Most NLP research uses this kind of tokenization technique, such as by [10] in semantic similarity, [4][9] in text classification, [11], [12] in information retrieval, [13], [14] in clustering, [15]–[17] in sentiment analysis, and much more.

Usually tokenization separates each word in a sentence as one token based on the spaces between words, but in fact, not all words in a sentence can be separated. There are words that must remain in pairs so that the meaning of the sentence remains correct. Separating a sentence into its constituent words can result in the meaning of a word deviating far from its actual context [18].

There are several publications that state that tokens are not just one word, but can be several words or even one sentence [10][13][14][19]. There is also research into finding multi-word expressions (MWE) or combinations of words that must be paired to make sense, such as by [20]–[23]. Most of this research was conducted for documents in English and other languages, including languages that do not recognize spaces as separators between words, such as Mandarin, Japanese or Thai. Research on Indonesian language texts is still limited. All the research above is only for finding word pairs and not for tokenization.

Methods that have been used in previous research include statistics, linguistic, dictionaries, and machine learning. The statistical method calculates the frequency of co-occurrence of two words. Linguistic methods match grammatical patterns based on the types of word labels. Searching for word pairs in the dictionary, that's the dictionary method. Machine learning methods use a set of datasets to predict the output.

Tokens consisting of several words are referred to as multi-word tokens. Multi-word tokens must be in the same sentence and same sentence element. In paragraphs that contain many sentences, it is necessary to segment the sentences so that each sentence is separated from each other. In order to segment a sentence, it is very important to know where the sentence boundaries are. It is not easy to find sentence boundaries because there is ambiguity from sentence boundary punctuation.

In Indonesian there are 5 sentence elements, namely Subject (Subjek), Predicate (Predikat), Object (Objek), Complement (Pelengkap) and Adverb (Keterangan) known as SPOK in Indonesia [24]. The subject and predicate elements must be present, while others may or may not be present. Each sentence element contains one or more words as word pairs.

Word pairs can only be formed in the same sentence element. Therefore, it is important to be able to perform sentence structure extraction. This is not taken into account by previous studies. By extracting the sentence structure, each sentence element can be treated as a token, at least for Subject and Object. This paper proposes a new method for sentence tokenization based on sentence structure in Indonesian. This new method of sentence tokenization will generate single-word and multi-word tokens simultaneously. That's our contribution. To our best knowledge, there is no research on this. This research uses Indonesian, but can be adapted to other languages that use spaces by customizing and retraining the dataset.

To find out the effectiveness of single-word and multi-word tokens, a sentence similarity test was carried out on both types of tokens. From the test results, it shows that multi-word tokens are able to determine word similarity better than single-word tokens.

This paper divided into several sections. In Section II, we review the related work on multi-word tokenization including multi-word expression, Section III, we give an overview of the proposed method including sentence segmentation, sentence structure extraction and dataset preparation. Section IV, we provide the result and discussion, and finally, Section V, concludes this paper.

## II. LITERATURE REVIEW

This paper is inseparable from the previous studies that have been conducted by researchers. The previous studies are summarized in this section, especially those related to multi-word tokenization. There are several methods used in previous research, such as statistics, linguistics, dictionary, and machine learning. We found two research in Indonesian language, that is [25] which perform 2-word extraction to obtain multi-word expression candidates by applying some rules and filtering using a dictionary. Researcher [26] also used rule-based methods and built two dictionaries (close class tagging and multi-word expression dictionary). This dictionary will store two or more words with POS tags of nouns, verbs and adjectives. The study [27] examines the tokenization process using a phrase detection-based approach.

Research in Serbian language with agricultural engineering domain conducted by [28] provides a hybrid approach by combining linguistic and statistical information. The Candidate terms are obtained using the frequency of occurrence of text sequences in the corpus. In an effort to obtain multi-word expressions, the author in [20] examined an implementation in Turkish used four methods: first, statistical methods to calculate high co-occurrence frequencies, second, linguistic methods through POS patterns, third, candidates from idiom dictionaries, and the last is specialized domains such as term dictionaries. Research that presents a method for identification of chemical terms as multi-words was conducted by [23]. In his research, the Multiword Identifying and Representing (MIR) method was implemented to recognize multi-word phrases in chemical literature with an unsupervised data-driven model and the identified phrases were added to the vocabulary. This research uses statistical and linguistic methods without expert annotations. Author in [29] created the MwTEXT architecture,

for automatic extraction of multi-word terms from unannotated computer science domain English documents. This method uses statistical, linguistic, and logic-based methods and hybrid techniques and focuses only on lexical patterns such as (N P N), (N P N + N), and (N P N P N).

The study [21] built a hybrid approach with the combination of Bi-LSTM + word correlation level and K-Means Clustering to detect MWEs for multiple languages without manual features. Author in [30] proposed a neural network model for learning fixed-size word representations from arbitrary chunks with word embedding. Implementation in French created MWE for Russian dictionary (RuThes). Multi-word expression recognition measure based on similarity of phrase distribution and word components is used for statistical and linguistic methods as well as for word embedding. Author in [31] focus on annotating different types of lexicalized and institutionalized phrases with main goal is to identify MWEs that are perceived as complex by readers and need to be simplified overall. A number of hand-crafted features form the basis for predicting MWE complexity.

From the previous research above, as far as we know, there is no research with a method based on sentence structure as proposed by this research.

## III. PROPOSED METHOD

The general tokenization process is shown in Fig. 1. This process works by receiving input in the form of sentences and identifying each word as a token by using spaces as separators between words, resulting in single word token. The number of tokens equals the number of words.

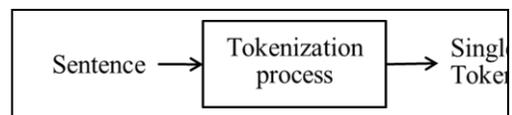


Fig. 1. General tokenization process.

This tokenization method is widely used, but it can also cause inaccuracies, such as:

1) The same word or token, will be considered to have the same meaning even if it is in a different order so that only one token will be used and the other tokens will be ignored [32]. Example :

Token in English : 'sakura', 'dewi', 'looks', 'at', 'sakura', 'tree', 'in', 'Japan.'

Token in Indonesian : 'sakura', 'dewi', 'memandang', 'pohon', 'sakura', 'di', 'Jepang'

The first token and the fifth token, will be considered to have the same meaning even though they are semantically different. One of them will be ignored.

2) When two or more words are combined and form a whole, a new meaning will be created that is different from each of the constituent words. Example :

Token in English: 'green table'

Token in Indonesian: 'meja hijau'

In Indonesian, 'meja hijau' means the court, a place to find the truth. If these two words are separated into 'meja' and 'hijau' then the meaning becomes different, the first is a piece of furniture that has a flat surface as a table top and legs as a support and the second is one of the base colors.

3) Not only the word meaning problem, but also the Part-of-Speech (POS) ambiguity problem. The POS of a single word token can vary. For example, separating the two words 'memberi makan' (in English: feeding), consists of the word 'memberi' with POS as the verb and the word 'makan' as the noun (since it is something that is given), but in other contexts such as 'kuda makan rumput', the POS of the word 'makan' is as a verb.

From the previous description, it is known that there are words that cannot be separated or must still be combined. Current tokenization methods does not accommodate this.

The main elements of the proposed tokenization framework are shown in Fig. 2. The framework has two stages, namely sentence segmentation and sentence structure extraction.

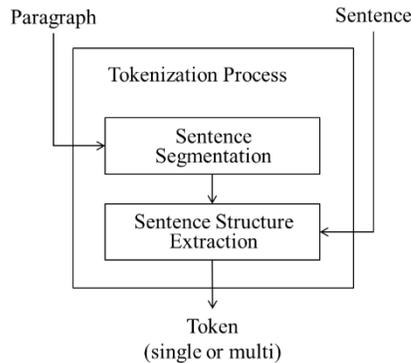


Fig. 2. The tokenization framework block diagram.

The input can be in the form of paragraphs or sentences. If the input is a paragraph, it will go through the sentence segmentation stage. This stage will split the paragraph into separate sentences. These sentences, whether they are new input or output from the first stage will be processed in the sentence structure extraction stage.

The output is a sentence structure with its elements (SPOK). Each sentence element is a token. In other words, sentence structure extraction is a tokenization process. These tokens are then used in natural language processing applications.

#### A. Sentence Segmentation

The task of sentence segmentation can be performed by detecting sentence boundaries [33]. The general pattern of a sentence is that it begins with a capital letter and ends with a special punctuation mark such as a period, question mark, or exclamation mark. The ability to recognize punctuation is a key requirement for knowing sentence boundaries to divide a paragraph into sentences. In this study the sentence segmentation process is described in Fig. 3.

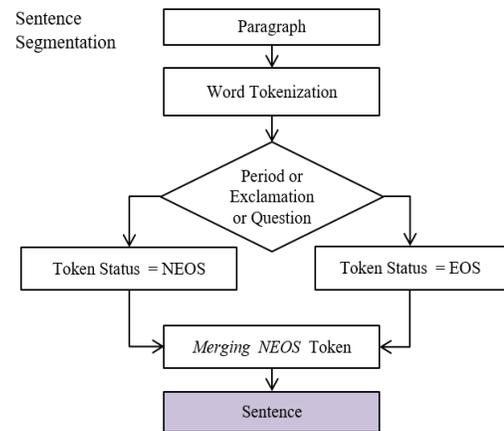


Fig. 3. Sentence segmentation diagram.

1) Word Tokenization, is a tokenization process as commonly used, breaking text data into words [8][34]. If there are punctuation marks then they will be attached to this token.

2) Punctuation Checking, is the process of checking the punctuation attached to the token, one of which is a period, question mark, or exclamation mark.

3) If the punctuation on the token is one of the three sentence-ending punctuation marks, the token will be assigned EOS status. Otherwise, it will be assigned NEOS status.

4) Combining NEOS Tokens. All NEOS will be combined into one sentence after finding EOS.

All tokens with NEOS status are combined into one new sentence and tokens with EOS status become the last word in the sentence. The next token will be the first word of the next sentence. This sentence will be used as input for the next process.

#### B. Sentence Structure Extraction

There are five sentence elements in Indonesian, namely Subject (Subjek), Predicate (Predikat), Object (Objek), Complement (pElengkap) and Adverb (Keterangan). Each sentence consists of at least a Subject and a Predicate and these two elements are arranged sequentially. The sentence elements Object, Complement and Adverb can be used or not used. The combination of these sentence elements forms a sentence structure pattern like SP, SPO, SPOK, SPOE, SPK, SPE, SPEK and SPOEK. Each word or words in each element of the sentence is a unit. Words or tokens that are in different sentence elements cannot be combined into one unit.

The sentence extraction process will identify sentence elements and classify each word in each sentence element.

This will facilitate the tokenization process, especially in determining multi-word tokens. The sentence structure extraction method in this study is as shown in the Fig. 4.

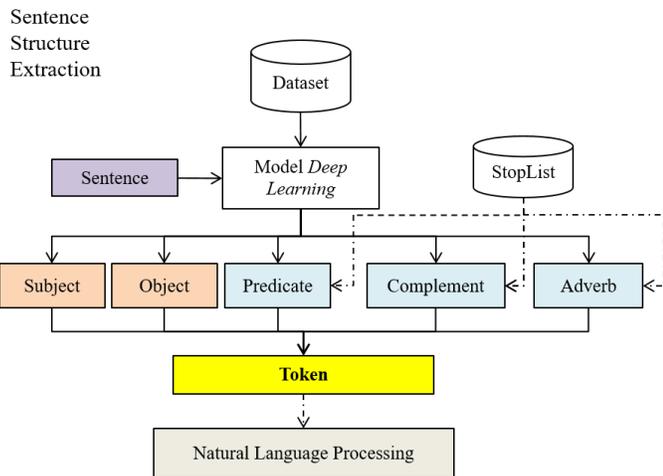


Fig. 4. Sentence structure extraction diagram.

The stages of the extraction process are as follows:

1) The process will accept input in the form of simple and active sentences.

2) A pre-trained deep learning model will predict sentence structure of the input sentence. The model has been trained using a dataset containing a collection of simple and active sentences in Indonesian, complete with labels. The embedded label is the identity of the sentence structure in the BIO tagging format. Label B (for "beginning") indicates as part of a multi-word token with position as the first word. The label I (for "inside") also indicates as part of a multi-word token with the position as the next word and the label O (for "outside") indicates as a stand-alone token or single word token. The dataset is in csv file format with an example as shown in Fig. 5.

This dataset contains 45,079 tokens from 4,740 sentences in Indonesian, with a minimum token range of 2 words and a maximum of 17 words per sentence. The distribution of each sentence element contained in the dataset is shown graphically in Fig. 6.

Fig. 5. Sentence structure dataset.

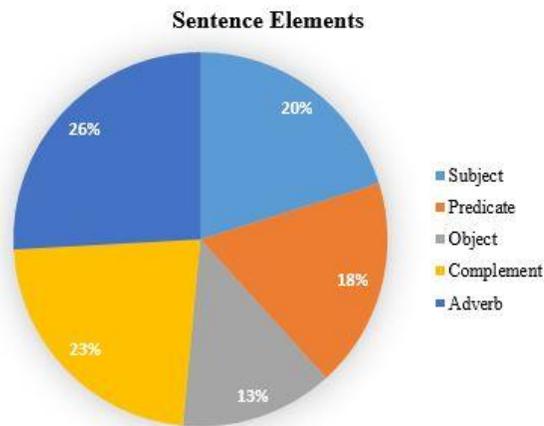


Fig. 6. Distribution of sentence elements in the dataset.

This dataset was trained using the pre-trained Bidirectional Encoder of Transformers (BERT) model. By dividing 80% as training data and 20% as test data and 10 epochs, an F1-score of 0.7 was obtained. These results show that the model and dataset that have been built are good enough, but need to be improved in the future.

3) The output of this process is a sentence structure prediction with sentence elements, namely Subject (SUB), Predicate (PRE), Object (OBJ), Complement (PEL), and Adverb (KET). There are nine types of adverbs in the dataset so there are thirteen sentence elements as listed in Table I.

Each token or word must be a member of one of the sentence elements. Each sentence element can consist of one or more than one word.

The output of the predicted sentence element will be written in the format of a BIO-tag label and the abbreviation of the sentence element, e.g. 'O-SUB' consists of the label O and the abbreviation SUB which means the word has no word pairs and with the Subject role.

TABLE I. SENTENCE ELEMENTS

No.	Sentence element		Abbreviation
1.	Subject	Subject	SUB
2.	Predicate	Predicate	PRE
3.	Object	Object	OBJ
4.	Complement	Complement	PEL
5.	Adverbs	Adverbs of time	KWK
6.		Adverbs of place	KTM
7.		Adverbs of purpose	KGU
8.		Adverbs of situation	KKD
9.		Adverbs of manner	KCR
10.		Adverbs of tools	KAL
11.		Adverbs of identity	KID
12.		Adverbs of participant	KPE
13.		Adverbs of condition	KSY

For sentence elements with more than one word, the first word will be labeled B ('beginning') and the remaining words will be labeled I ('inside' in BIO tags), e.g. 'B-SUB', 'I-SUB', 'I-SUB' which means there are three words that have the role of Subject and as one unit or one token. Such tokens are referred to as multi-word tokens. These tokens are then used in the NLP process.

#### IV. RESULT AND DISCUSSION

The experimental results of the proposed tokenization framework are quite good. In this section, the output will be discussed and sentence similarity tests will be conducted based on single word tokens and multi-word tokens.

##### A. The Output

As mentioned earlier, the outputs of this tokenization framework are sentence structures and sentence elements. Each sentence element can consist of a single word called a single-word token or multiple words called a multi-word token. One word means one token, multiple words also means one token. The number of sentence elements indicates the minimum number of tokens. Table II shows an example.

The first sentence consists of two words, the prediction results show that the first word is the Subject (O-SUB) and the second word is the Predicate (O-PRE). Both are independent because they are labeled O. Then each word is a single word token.

The second sentence consists of seven words. The first word 'Tim' is labeled 'B-SUB' and the second word 'Argentina' is labeled 'I-SUB' which indicates that both are in the same group which is Subject (SUB). So both should remain as one with the meaning of a group of soccer players from Argentina. Separating the two words will lose the original meaning. That is, the Subject is a combination of the words 'Tim' and 'Argentina' to become 'Tim Argentina'. This is a multi-word token.

Likewise, the fourth to seventh words are adverb groups (KTM), so these four words are a single unit. In this second sentence, there is also a word labeled 'O-PRE', namely 'win'. This means that the word 'win' has the role of a Predicate that stands alone, and is a single word token.

Therefore, it can be seen that the second sentence only has three tokens for Subject, Predicate, and Adverb. More details in Table III.

In the third sentence, there are three groups of sentence elements consisting of more than one word, namely words labeled Predicate (PRE), Object (OBJ), and Complement (PEL). Only the subject (SUB) stands alone because it is labeled O. The complete information can be seen in Table IV.

In Table IV, it is clear that the Subject is a one-word token 'Prajurit', the Predicate is a multi-word token 'mulai memasuki', on the Object there are two words 'area pertempuran' as multi-word tokens and the Complement consists of three words 'dengan senjata lengkap' as multi-word tokens.

##### B. Sentence Elements as Token

As explained earlier, a sentence element can be a token. A sentence extraction result that produces three sentence elements means it has three tokens. A sentence will have at least two tokens. Tokens can be single-word tokens or multi-word tokens.

TABLE II. INPUT AND PREDICTION OF SENTENCE ELEMENTS

No	Lang	Input Sentence	Output Prediction	
			Tokens	Sentence Elements
1.	INA	Amir mandi	['Amir', 'mandi']	['O-SUB', 'O-PRE']
	EN	Amir takes a bath		
2.	INA	Tim Argentina menang di Piala Dunia 2022	['Tim', 'Argentina', 'menang', 'di', 'Piala', 'Dunia', '2022']	['B-SUB', 'I-SUB', 'O-PRE', 'B-KTM', 'I-KTM', 'I-KTM']
	EN	The Argentina team won in the 2022 World Cup		
3.	INA	Prajurit mulai memasuki area pertempuran dengan senjata lengkap.	['Prajurit', 'mulai', 'memasuki', 'area', 'pertempuran', 'dengan', 'senjata', 'lengkap']	['O-SUB', 'B-PRE', 'I-PRE', 'B-OBJ', 'I-OBJ', 'B-PEL', 'I-PEL', 'I-PEL']
	EN	Soldiers began to enter the battle area with full weapons.		

TABLE III. SENTENCE STRUCTURE FOR EXAMPLE NO. 2

Source	Tim Argentina menang di Piala Dunia 2022						
Initial Token	Tim	Argentina	menang	di	Piala	Dunia	2022
Output Labels	B-SUB	I-SUB	O-PRE	B-KTM	I-KTM	I-KTM	I-KTM
Sentence Elements	Subject		Predicate	Adverb of Place			
Proposed Token	'Tim Argentina'		'menang'	'di Piala Dunia 2022'			
	Multi-word		Single word	Multi-word			

TABLE IV. SENTENCE STRUCTURE FOR EXAMPLE NO. 3

Source	Prajurit mulai memasuki area pertempuran dengan senjata lengkap.							
Initial Token	Prajurit	mula i	memasu ki	area	pertempura n	denga n	senjat a	lengka p
Output Labels	O-SUB	B-PRE	I-PRE	B-OBJ	I-OBJ	B-PEL	I-PEL	I-PEL
Sentence Element	Subject	Predicate		Object		Complement		
Proposed Token	'Prajurit'	'mulai memasuki'		'area pertempuran'		'dengan senjata lengkap'		
	Single word	Multi-word		Multi-word		Multi-word		

However, not all multi-word tokens derived from sentence elements can be assigned as end tokens. The contents of multi-word tokens can be words that do not provide important information.

In the second sentence above, there is the word 'di' in the adverb of place with a multi-word token. The multi-word token

in the third sentence contains the word 'mulai' in the Predicate and the word 'dengan' in the Complement. These words can be ignored and have no effect on the token. Such words are known as stopwords.

From the example sentences above, stopwords can appear in Predicate, Complement, or Adverb. There are almost no stopwords in Subject and Object. Therefore, multi-word tokens in Predicate, Complement, and Adverb need to be filtered first. These unnecessary words will be removed before providing tokens. Filtering is done by comparing the contents of the multi-word tokens of the three sentence elements with a database containing words that fall into the category of stopwords.

### C. Evaluation

The outputs of this framework are single word tokens and multi-word tokens. To get an overview of the two types of tokens, the following is an evaluation of both in determining sentence similarity.

The evaluation is done using the token lexical similarity method. Overlap Coefficient, Jaccards Index, Jaccards Distance, Dice Coefficient and Cosine Similarity methods will be used for single word tokens, while Dice-Index Coefficient for multi-word tokens.

Some of the stages of evaluation are as follows:

1) Defines a set of single-word tokens and multiple-word tokens in sentences.

2) Perform statistical calculations:

a) For single word token.

- Counts the number of tokens in the sentence, which is mathematically symbolized as  $|K_1|$ .
- Counts the number of tokens that appear in both sentences, symbolized as  $|K_1 \cap K_2|$ .
- Counts the number of tokens derived from the two sentences, and is symbolized as  $|K_1 \cup K_2|$ .

b) For multi-word tokens:

- Counts the core (head) token on each token, symbolized as  $|h_1|$  and  $|h_2|$ . Head is a word whose meaning is included in the meaning of another word.
- Perform token combinations according to the token order.
- Counts the number of core tokens (head) present in both multi-word tokens, symbolized as  $|h_1 \cap h_2|$ .
- Sum the core (head) tokens, symbolized by  $|h_1| + |h_2|$ .
- Counts the number of tokens present in both multi-word tokens and is symbolized as  $|M_1 \cap M_2|$ .
- Counts the number of tokens from both multi-word tokens, symbolized as  $|M_1| + |M_2|$ .

c) Measuring sentence similarity

Measuring the similarity between sentence1 and sentence2 basically determines how many similarity tokens there are in each sentence divided by the normalization factor.

The sentence similarity measurement function used is as follows:

- Overlap Coefficient: is the size of the overlap of the sets  $K_1$  and the sets  $K_2$  divided by the smallest size between  $K_1$  and  $K_2$ .

$$OC(K_1, K_2) = \frac{|K_1 \cap K_2|}{\min(|K_1|, |K_2|)} \quad (1)$$

- Jaccard Index: is the Intersection over Union size of the sets  $K_1$  and  $K_2$ .

$$JI(K_1, K_2) = \frac{|K_1 \cap K_2|}{|K_1 \cup K_2|} = \frac{|K_1 \cap K_2|}{|K_1| + |K_2| - |K_1 \cap K_2|} \quad (2)$$

- Jaccard Distance: Measures the degree of difference of the two sets, or by subtracting 100% with the Jaccard Index.

$$JD(K_1, K_2) = 1 - JI(K_1, K_2) = \frac{|K_1 \cup K_2| - |K_1 \cap K_2|}{|K_1 \cup K_2|} \quad (3)$$

- Dice Coefficient: measures two times the number of tokens shared in both sentences divided by the total number of tokens in both sentences.

$$DC(K_1, K_2) = \frac{2|K_1 \cap K_2|}{|K_1| + |K_2|} \quad (4)$$

- Cosine Similarity, with the formula:

$$CS(K_1, K_2) = \frac{|K_1 \cap K_2|}{\sqrt{|K_1| \cdot |K_2|}} \quad (5)$$

The following three sentences are used as test data.

1)  $K_1$  = “walikota solo memberikan apresiasi kepada Agnes.” (The mayor of solo city gave his appreciation to Agnes.)

2)  $K_2$  = “agnes monica adalah penyanyi solo wanita berbakat.” (agnes monica is a talented female solo singer.)

3)  $K_3$  = “pemerintah kota solo mendapat hibah dari pangeran arab saudi.” (she solo city government received a grant from the prince of saudi arabia.)

By using the formula described above, the calculation results are as follows in Table V. From the table, it can be concluded that the first sentence is more similar to the second sentence.

Meanwhile, the proposed tokenization process generates tokens according to the sentence structure as follows:

For  $K_1$ , S=“walikota solo”, P=“memberikan”, O=“apresiasi”, C=“kepada agnes”.

For  $K_2$ , S=“agnes monica”, P=“adalah”, O=“penyanyi solo wanita”, A=“berbakat”.

TABLE V. SENTENCE SIMILARITY FOR SINGLE WORD TOKEN

	$K_1$	$K_2$	$K_3$	$K_1, K_2$	$K_1, K_3$	$K_2, K_3$
$ K_n $	6	7	9			
$ K_x \cap K_y $				2	1	1
$ K_x \cup K_y $				11	14	15

Overlap Coefficient	0.3333	0.1667	0.1429
Jaccard Index	0.1818	0.0769	0.0714
Jaccard Distance	0.8182	0.9231	0.9286
Dice Coefficient	0.3077	0.1429	0.1333
Cosine Similarity	0.3086	0.1443	0.1336

For  $K_3$ ,  $S$ ="pemerintah kota solo",  $P$ ="mendapat",  $O$ ="hibah",  $A$ ="dari pangeran arab saudi".

Multi-word token similarity measurement uses the concept of lexical similarity based on identifying the common sequence of each token. It is based on the hypothesis that the head is a hyponym of the same term, which is denoted as  $h_n$ . The visualization of the hyponyms of the multi-word tokens in the above three sentences is shown in the Fig. 7 below.

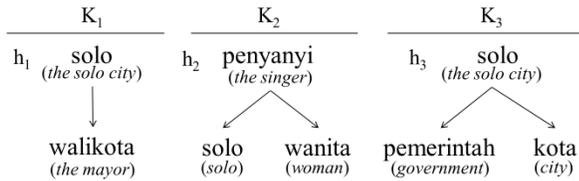


Fig. 7. Hyponyms referring heads.

The word sequence of the multiword token  $P(t)$  references the set of all sequences in  $t$ . The lexical similarity between multi-word tokens  $t_1$  and  $t_2$  is measured based on the Dice-like coefficient formula as follows:

$$KMK(M_1, M_2) = \frac{|P(h_1) \cap P(h_2)|}{|P(h_1)| + |P(h_2)|} = \frac{|P(t_1) \cap P(t_2)|}{|P(t_1)| + |P(t_2)|} \quad (6)$$

The numerator in the formula indicates the set of shared constituents (constituents present in both tokens), while the denominator refers to the total number of constituents.

The multi-word token obtained from sentence structure extraction are shown in Table VI.

TABLE VI. MULTI-WORD TOKEN

Stc	Multi-word Token	Core Token (head)		Constituent order	
		$P(h)$	$ P(h) $	$P(t)$	$ P(t) $
$K_1$	'walikota solo'	'solo'	1	{walikota, solo, walikota solo}	3
$K_2$	'agnes monica'	-	-	-	-
	'penyanyi solo wanita'	'penyanyi'	1	{penyanyi, solo, wanita, penyanyi solo, solo wanita, penyanyi solo wanita}	6
$K_3$	'pemerintah kota solo'	'solo'	1	{pemerintah, kota, solo, pemerintah kota, kota solo, pemerintah kota solo}	6

TABLE VII. SENTENCE SIMILARITY LEVEL FOR MULTIWORD TOKEN

Formula	Description	$K_1, K_2$	$K_1, K_3$	$K_2, K_3$
$ P(h_x) \cap P(h_y) $	The sum of the same terms in both heads	0	1	0
$ P(h_x)  +  P(h_y) $	The total number of terms on each head	2	2	2

$ P(t_x) \cap P(t_y) $	The sum of the same terms in both constituents	1	1	1
$ P(t_x)  +  P(t_y) $	The total number of terms on each constituents	9	9	12
Similarity		0.11	0.61	0.08

By using the Dice-like coefficient formula, the level of similarity of multiword tokens is obtained as shown in Table VII.

From the table above, the multi-word tokens in the first sentence are similar to the third sentence compared to the second sentence, and the multi-word tokens in the second sentence are very different from the third sentence.

From the similarity measurement of the two sentences above, there is a difference in results between single word tokens and multi-word tokens. The measurement with single word tokens concludes that the first sentence and the second sentence are more similar than the other sentences.

While the measurement with multi-word tokens states that the first sentence and the third sentence are more similar than the first and second sentences. Both have the same measurement result, that the second and third sentences are least similar.

In human judgment, the first and third sentences are similar, just like the measurement results of multi-word tokens. This shows that multi-word tokens also have advantages and can help NLP work.

#### D. Performace

To evaluate the quality of the proposed method, we conducted a manual evaluation of 100 sentences. The evaluation was done by checking the supposed multi-word tokens and then compared with the multi-word tokens extracted by the proposed method, with the results as shown in the Table VIII.

From the table, we can calculate Precision and Recall using the following formula:

$$P = \frac{\text{Correctly extracted multi-word tokens}}{\text{Total extracted multi-word tokens}} \quad (7)$$

$$R = \frac{\text{Correctly extracted multi-word tokens}}{\text{Multi-word tokens should be}} \quad (8)$$

TABLE VIII. EXTRACTED MULTI-WORD TOKEN

Number of sentences	Number of tokens	Correctly extracted sentence structure	Extracted		
			Correctly extracted multi-word tokens	Total extracted multi-word tokens	Multi-word tokens should be
100	709	84	204	221	237

And the results are  $P = 0.92$  and  $R = 0.86$ . The success of extracting multi-word tokens correctly is quite dominant, out of 221 multi-word tokens extracted, 204 of them are correct. While the R value has a value of 0.86 which is obtained from 204 correct multi-word tokens out of 237 multi-word tokens that can be generated. These results provide information that

the proposed method is able to extract sentence structure and at the same time produce multi-word tokens that are quite accurate.

We also conducted a comparison with three other studies on multi-word tokens or similar from [21], [27] and [29]. Methods used by [21] are hybrid to train a multi-word expression detector for multiple languages without any manually encoded features. The methods used by [27] is a rule-based. The methods used by [29] are statistical, linguistic and logic-based methods and hybrid techniques, for the automatic extraction of multi-word terms from unannotated computer science domain English documents.

A comparison between these four methods is shown in Table IX.

TABLE IX. METHOD COMPARISON

Liang et al. [21]	Putranto et al. [27]	Thanawala et al. [29]	Propose Method
a hybrid approach, which combines Bi-directional LSTM (Bi-LSTM), phrase head word expansion and cluster to identify three types of multi-word expression	a rule that contains combinations of word classes that are most likely to form phrases.	Using shallow parsing and syntactic structure analysis and using a rule-based linguistic approach pattern.	Sentence structure extraction
compound nouns, verb construction and idiom.	Verbal, Nominal, Adverbial, Pronominal, Adjectival phrase rule.	Lexical patterns such as (N P N), (N P N + N), (N P N P N).	Dataset model
Sequence features, word correlation degree and three types of multi-word expression	The classification model obtained is more optimal.	Output in various forms of noun phrases	Output in form of sentence element
Precision = 0.92	0.79	0.87	0.92
Recall = 0.92	0.84	1.00	0.86

Each method has advantages and disadvantages. However, by preparing and training the sentence structure dataset, the proposed method is excellent in predicting the sentence structure elements. Each element is a token, either a single token or a multi-word token. Thus, this method does not rely on manually constructed lexical patterns. The method is highly adaptable and evolves as new data becomes available.

## V. CONCLUSION

A tokenization process that generates single-word tokens and multi-word tokens simultaneously is possible. This is proposed through this research. To our knowledge, we are the first to propose this tokenization method based on sentence structure, which is expected to inspire new research with new ideas. Providing a complete dataset is a very important factor for successful sentence structure prediction. The predicted sentence element (SPOK) can consist of one or more words, i.e. tokens. Multi-word tokens are more accurate than single-word tokens in terms of sentence similarity.

Multi-word tokens are worthy of further research. In the future, we will enhance the dataset with passive sentences and

also apply this approach for use in other types of cases such as NER.

## REFERENCES

- [1] El Haddadi, A. Fennan, A. El Haddadi, Z. Boulouard, and L. Koutti, "Mining unstructured data for a competitive intelligence system XEW," SIIE 2015 - 6th Int. Conf. "Information Syst. Econ. Intell., pp. 146–149, 2015, doi: 10.1109/ISEI.2015.7358737.
- [2] R. Talib, M. Kashif, S. Ayesha, and F. Fatima, "Text Mining: Techniques, Applications and Issues," Int. J. Adv. Comput. Sci. Appl., vol. 7, no. 11, pp. 414–418, 2016, doi: 10.14569/ijacsa.2016.071153.
- [3] M. M. Samia, A. Rajee, R. Hasan, M. O. Faruq, and P. C. Paul, "Aspect-based Sentiment Analysis for Bengali Text using Bidirectional Encoder Representations from Transformers (BERT)," vol. 13, no. 12, 2022.
- [4] H. X. Huynh, L. X. Dang, N. Duong-Trung, and C. T. Phan, "Vietnamese Short Text Classification via Distributed Computation," Int. J. Adv. Comput. Sci. Appl., vol. 12, no. 7, pp. 23–31, 2021, doi: 10.14569/IJACSA.2021.0120703.
- [5] M. Allahyari et al., "A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques," arXiv, no. July, 2017.
- [6] R. A. Farouk, M. H. Khafagy, M. Ali, K. Munir, and R. M. Badry, "Arabic Semantic Similarity Approach for Farmers' Complaints," Int. J. Adv. Comput. Sci. Appl., vol. 12, no. 10, pp. 348–358, 2021, doi: 10.14569/IJACSA.2021.0121038.
- [7] W. L. Roldan-Baluis, N. A. Zapata, and M. S. M. Vásquez, "The Effect of Natural Language Processing on the Analysis of Unstructured Text: A Systematic Review," Int. J. Adv. Comput. Sci. Appl., vol. 13, no. 5, pp. 43–51, 2022, doi: 10.14569/IJACSA.2022.0130507.
- [8] N. M. Ibrahim, W. M. S. Yafooz, A. M. Emara, and A. Abdel-wahab, "Utilizing Deep Learning in Arabic Text Classification Sentiment Analysis of Twitter," vol. 13, no. 12, pp. 830–838, 2022.
- [9] M. Usman, Z. Shafique, S. Ayub, and K. Malik, "Urdu Text Classification using Majority Voting," Int. J. Adv. Comput. Sci. Appl., vol. 7, no. 8, pp. 265–273, 2016, doi: 10.14569/ijacsa.2016.070836.
- [10] I. A. Norabid and F. Fauzi, "Rule-based Text Extraction for Multimodal Knowledge Graph," Int. J. Adv. Comput. Sci. Appl., vol. 13, no. 5, pp. 295–304, 2022, doi: 10.14569/IJACSA.2022.0130535.
- [11] M. Krallinger, O. Rabal, A. Lourenço, J. Oyarzabal, and A. Valencia, "Information retrieval and text mining technologies for chemistry," Chem. Rev., vol. 117, no. 12, pp. 7673–7761, 2017, doi: 10.1021/acs.chemrev.6b00851.
- [12] S. P. Panda, V. Behera, A. Pradhan, and A. Mohanty, "A Rule-based Information Extraction System," Int. J. Innov. Technol. Explor. Eng., vol. 8, no. 9, pp. 1613–1617, 2019, doi: 10.35940/ijitee.i8156.078919.
- [13] J. Joseph and J. R. Jeba, "Information Extraction using Tokenization and Clustering Methods," Int. J. Recent Technol. Eng., vol. 8, no. 4, pp. 3690–3692, 2019, doi: 10.35940/ijrte.d7943.118419.
- [14] S. A. Fahad, "Design and Develop Semantic Textual Document Clustering Model," J. Comput. Sci. Inf. Technol., vol. 5, no. 2, pp. 26–39, 2017, doi: 10.15640/jcsit.v5n2a4.
- [15] H. Juwiantho, E. I. Setiawan, J. Santoso, and M. H. Purnomo, "Sentiment Analysis Twitter Bahasa Indonesia Berbasis Word2Vec Menggunakan Deep Convolutional Neural Network," J. Teknol. Inf. dan Ilmu Komput., vol. 7, no. 1, pp. 181–188, 2020, doi: 10.25126/jtiik.202071758.
- [16] E. W. Pamungkas and D. G. P. Putri, "An experimental study of lexicon-based sentiment analysis on Bahasa Indonesia," Proc. - 2016 6th Int. Annu. Eng. Semin. Ina. 2016, pp. 28–31, 2017, doi: 10.1109/INAES.2016.7821901.
- [17] H. Sudira, A. L. Diar, and Y. Ruldeviyani, "Instagram Sentiment Analysis with Naive Bayes and KNN: Exploring Customer Satisfaction of Digital Payment Services in Indonesia," 2019 Int. Work. Big Data Inf. Secur. IWBIS 2019, pp. 21–26, 2019, doi: 10.1109/IWBIS.2019.8935700.
- [18] A. Hamzah, A. Susanto, F. Soesianto, and J. E. Istyanto, "Perbandingan Feature Kata dan Frasa dalam kinerja Clustering dokumen teks berbahasa Indonesia," in Seminar Nasional Aplikasi Teknologi Informasi (SNATI), 2007, no. SNATI, p. B-53-B-58.

- [19] U. Rahardja, T. Hariguna, and W. M. Baihaqi, "Opinion mining on e-commerce data using sentiment analysis and k-medoid clustering," *Proc. - 2019 12th Int. Conf. Ubi-Media Comput. Ubi-Media 2019*, pp. 168–170, 2019, doi: 10.1109/Ubi-Media.2019.00040.
- [20] S. K. Metin and M. Taze, "A procedure to build multiword expression data set," *2nd Int. Conf. Comput. Commun. Syst. ICCCS 2017*, pp. 46–49, 2017, doi: 10.1109/CCOMS.2017.8075264.
- [21] Y. Liang, H. Tan, H. Li, Z. Wang, and W. Gui, "A language-independent hybrid approach for multi-word expression extraction," *Proc. Int. Jt. Conf. Neural Networks*, vol. 2017-May, pp. 3273–3279, 2017, doi: 10.1109/IJCNN.2017.7966266.
- [22] S. Agrawal, R. Sanyal, and S. Sanyal, "Hybrid method for automatic extraction of multiword expressions," *Int. J. Eng. Technol.*, vol. 7, no. 2, pp. 33–38, 2018, doi: 10.14419/ijet.v7i2.6.10063.
- [23] L. Huang and C. Ling, "Representing Multiword Chemical Terms through Phrase-Level Preprocessing and Word Embedding," *ACS Omega*, vol. 4, no. 20, pp. 18510–18519, 2019, doi: 10.1021/acsomega.9b02060.
- [24] D. Gunawan, H. P. Siregar, and O. Salim Sitompul, "Identifying Sentence Structure in Bahasa Indonesia by Using POS Tag and LALR Parser," *5th Int. Conf. Comput. Eng. Des. ICCED 2019*, 2019, doi: 10.1109/ICCED46541.2019.9161125.
- [25] D. Gunawan, A. Amalia, and I. Charisma, "Automatic extraction of multiword expression candidates for Indonesian language," *Proc. - 6th IEEE Int. Conf. Control Syst. Comput. Eng. ICCSCE 2016*, no. November, pp. 304–309, 2017, doi: 10.1109/ICCSCE.2016.7893589.
- [26] F. Rashel, A. Luthfi, A. Dinakaramani, and R. Manurung, "Building an Indonesian rule-based part-of-speech tagger," *Proc. Int. Conf. Asian Lang. Process. 2014, IALP 2014*, pp. 70–73, 2014, doi: 10.1109/IALP.2014.6973521.
- [27] H. A. Putranto, O. Setyawati, and W. Wijono, "Effect of Phrase Detection with POS-Tagger on Sentiment Classification Accuracy using SVM," *J. Nas. Tek. Elektro dan Teknol. Inf.*, vol. 5, no. 4, pp. 252–259, 2016, doi: 10.22146/jnteti.v5i4.271.
- [28] V. Pajić, S. Vujičić Stanković, R. Stanković, and M. Pajić, "Semi-automatic extraction of multiword terms from domain-specific corpora," *Electron. Libr.*, vol. 36, no. 3, pp. 550–567, 2018, doi: 10.1108/EL-06-2017-0128.
- [29] P. Thanawala and J. Pareek, "MwTExt: automatic extraction of multiword terms to generate compound concepts within ontology," *Int. J. Inf. Technol.*, vol. 10, no. 3, pp. 303–311, 2018, doi: 10.1007/s41870-018-0111-6.
- [30] J. Legrand and R. Collobert, "Phrase Representations for Multiword Expressions," *Proc. 12th Work. Multiword Expressions*, no. 2011, pp. 67–71, 2016, doi: 10.18653/v1/w16-1810.
- [31] E. Kochmar, S. Gooding, and M. Shardlow, "Detecting multiword expression type helps lexical complexity assessment," *Lr. 2020 - 12th Int. Conf. Lang. Resour. Eval. Conf. Proc.*, no. May 2020, pp. 4426–4435, 2020.
- [32] J. Rabelo, M. Y. Kim, and R. Goebel, "Combining similarity and transformer methods for case law entailment," *Proc. 17th Int. Conf. Artif. Intell. Law, ICAIL 2019*, pp. 290–296, 2019, doi: 10.1145/3322640.3326741.
- [33] K. Lim and J. Park, "Real-world sentence boundary detection using multitask learning: A case study on French," *Nat. Lang. Eng.*, pp. 1–21, 2022, doi: 10.1017/S1351324922000134.
- [34] A. A. Jalal and B. H. Ali, "Text documents clustering using data mining techniques," *Int. J. Electr. Comput. Eng.*, vol. 11, no. 1, pp. 664–670, 2021, doi: 10.11591/ijece.v11i1.pp664-670.

# An Efficient Real-Time Weed Detection Technique using YOLOv7

Ch. Lakshmi Narayana<sup>1</sup>, Kondapalli Venkata Ramana<sup>2</sup>

Research Scholar, Department of CS&SE-Andhra University College of Engineering, Andhra University, Visakhapatnam, India<sup>1</sup>  
Associate Professor, Department of CS&SE-Andhra University College of Engineering, Andhra University, Visakhapatnam,  
India<sup>2</sup>

**Abstract**—Since farming is becoming increasingly more expensive, efficient farming entails doing so without suffering any losses, which is what the current situation desires. Weeds are a key issue in agriculture since they contribute significantly to agricultural losses. To control the weed, pesticides are now evenly applied across the entire area. This approach not only costs a lot of money but also harms the environment and people's health. Therefore, spot spray requires an automatic system. When a deep learning embedded system is used to operate a drone, herbicides can be sprayed in the desired location. With the continuous advancement of object identification technology, the YOLO family of algorithms with extremely high precision and speed has been applied in a variety of scene detection applications. We propose a YOLOv7-based object detection approach for creating a weed detection system. Finally, we used the YOLOv7 model with different parameters for training and testing analyzed on the early crop weed dataset and 4weed dataset. Experimental results revealed that the YOLOv7 model achieved the mAP@0.50, f1score, Precision, and Recall values for the bounding boxes as 99.6, 97.6, 99.8, and 95.5 respectively on the early crop weed dataset and 78.53, 79.83, 86.34, and 74.24 on 4weed dataset. The Agriculture business can benefit from using the suggested YOLOv7 model with high accuracy in terms of productivity, efficiency, and time.

**Keywords**—Weed detection; YOLOv7; early crop weed; deep learning

## I. INTRODUCTION

As of right now, losses from pests, diseases, and weeds can account for up to 40% of annual crop yields worldwide. In the years to come, this proportion is anticipated to rise sharply. Currently, the principal method of weeding in fields is to spray herbicides across a huge region. Leaving pesticide residues in the soil, this practice not only wastes resources but also pollutes the environment. As a consequence, precision spraying [1][2] effectively controls the growth of weeds in a field while using less pesticide, improving utilization, and avoiding chemical residue.

Quick and accurate weed detection in crop fields is crucial because it may serve as a foundation for the development of precision spraying systems. Many researches have been done so far on image-based techniques for the automated identification and categorization of weeds. For the purpose of enhancing weed detection accuracy in rice fields, [3] retrieved 101-dimensional characteristics from a picture of a weed, including color, shape, and texture. They achieved a recognition rate of 91.13 percent using deep belief networks

with fusion features. Two different classification techniques are presented by [4] to identify weed density in photos. Based on the grey level co-occurrence matrix (GLCM), the first approach used a Support Vector Machine (SVM) to get an accuracy of 73 percent, while the second method combined a Random Forest classifier with invariant scale and rotation moment features to achieve an accuracy of 86 percent. These methods have the drawback of not being effective against sedges and wide-leaf weeds. The artificial neural network (ANN) employed by [5] to identify various types of weeds was optimized using the bee algorithm (BA), and the ANN-BA attained an accuracy of 88.74 percent for the right channel and 87.96 percent for the left channel. The techniques utilized in the aforementioned research were aimed at enhancing recognition in conventional machine vision. Due to the minimal hardware requirements for operation, they are well-suited for practical deployment. However, the majority of these techniques only tested the effectiveness on low-density samples. The difficulties of opacity, clumping, light change and other natural environment characteristics are challenging to overcome.

Deep learning has been used to address weed detection issues in agriculture. Researchers have had success using various deep learning models for this task. In [6], employed the small YOLO-v3 for real-time application in a field of strawberry and tomato plants and succeeded in detecting goose grass with an accuracy of 82 percent. With the use of pre-trained Faster R-CNN, [7] achieved 65 percent accuracy, 68 recall, 66 F1 score, and 0.21 s inference time in recognizing late-season weed in soybean fields. The author [8] used Inception-ResNet-v2 as the basis and achieved F1 scores of 72.7 percent (at IoUall) and 96.9 percent for identifying agricultural plants and weeds (at IoU0.5). The study [9] used the Mask R-CNN to accurately extract weed from the "cranesbill seedling dataset." In [10] the author categorized the weed *Rumex obtusifolius* with a VGG-16 classification accuracy of 92.1%. The research [11] discovered that VGG-19, which had been tweaked to generate binary output, had the greatest classification accuracy of 98.7 percent for detecting volunteer potatoes in sugar beet in a comparison of Inception-v3 with AlexNet, VGG-19, GoogLeNet, ResNet-50, and ResNet-101.

Convolutional neural networks have been used by some researchers in recent years to try to detect weeds in rice fields. Fully convolutional networks were utilized by [12] to classify pixels in high resolution unmanned aerial vehicle (UAV)

imagery taken from a rice field (FCN). Their method had an accuracy rate of 91.96 percent and an average mean intersection over union (mean) of 84.73 percent. Using a semantic segmentation model called SegNet, IoU. In [13], detected the pixels in the image that corresponded to rice seedlings, weeds, and the backdrop. They can address the category imbalance by finding the class weight coefficients. Their approach has a greater accuracy of 92.7 percent when compared to FCN and U-Net.

However, the bulk of important research has only been able to recognize the leaves of certain plants, rather than actual photographs with intricate backgrounds in real settings. The techniques have poor stability and accuracy when applied to identify weeds in rice fields [15].

Large-scale weed picture collections must be carefully curated to create high-performing weed identification algorithms. Images of weed may be captured on a variety of platforms [16], incorporating field robots [18], portable camera sensors, and unmanned aerial vehicles (UAV) [17]. DeepWeeds [19], Early crop weed dataset [21], Open Plant Phenotype Database [22], and Dataset of food crops and weeds [23] are only a few examples from a recent assessment of 19 publicly accessible datasets for weed identification and plant recognition, published in [20]. These datasets are all made up of RGB (red-green-blue) photos. Currently, a large number of researches have demonstrated the effectiveness of deep learning object detectors in weed identification. These studies include those using the YOLO series, Faster R-CNN, Mask R-CNN, RetinaNet, and EfficientDet.

The YOLOv7 approach had been used in this study to address this issue and significantly enhance the performance for weed detection in the early weed dataset [21] and to assess the performance of a newly formed 4weed dataset [14] that has had no machine learning models applied to it up to this point. Finally, studies show that the YOLOv7 proposed in this study may successfully handle the problems related to weed identification in crops, achieving high accuracy and outstanding efficiency.

## II. METHODOLOGY

To build a framework for weed identification, we must finish data collection, model training, and multi-class plant species classification. Two main datasets were used in this study: The Early Crop Weeds dataset and the 4weed dataset. The dataset contains photos with varied resolutions that were translated into the same dimensions using the deep learning model input layer. After creating a suitable dataset, the gathered data is separated into 90% training and 10% testing sets. YOLOv7 is then trained for agricultural weed detection utilizing those data. The performance of the trained model is evaluated using multiple parameters. Fig. 1 depicts the proposed approach used for weed identification.

### A. Dataset

The early crop weed detection dataset contains 308 images that are taken from the early crop weed classification dataset [21] and the objects of interest are annotated with bounding boxes. This dataset contains 308 RGB images of four species at early growth stages. It includes images of 25 cotton, 67 velvet,

121 tomato, and 95 nightshade. Fig. 2 shows sample instances of the early crop weed detection dataset images.

The 4Weed [14] collection includes 618 RGB photos in total, which were collected at Purdue University under challenging field circumstances. The collection includes photos of four weed species that are often seen in corn and soybean production systems: Giant Ragweed, Foxtail, Cocklebur, and Redroot Pigweed Fig. 3.

The final dataset included 150 Giant Ragweed photos, 170 Redroot Pigweed images, 35 Cocklebur images, and 73 Foxtail images. You may get the dataset at <https://osf.io/w9v3j/>.

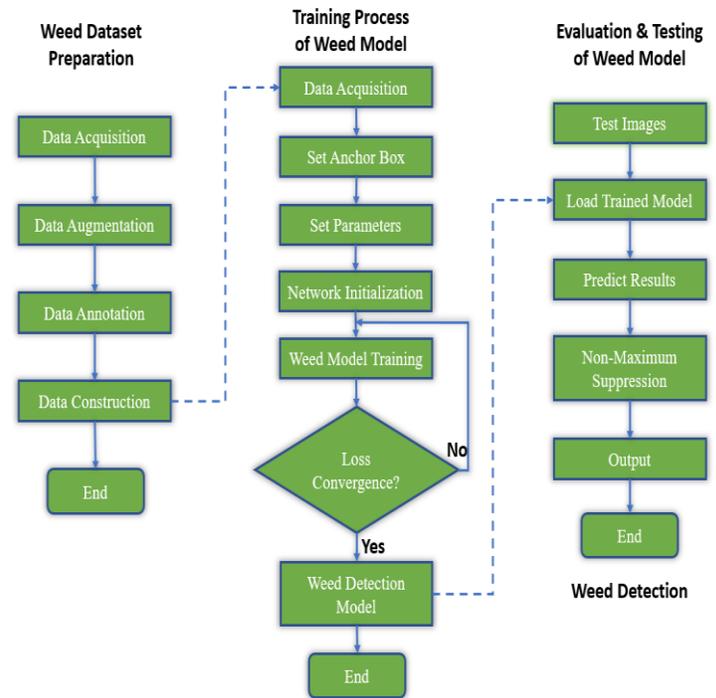


Fig. 1. Weed detection model flow chart for dataset, training, and detection process.

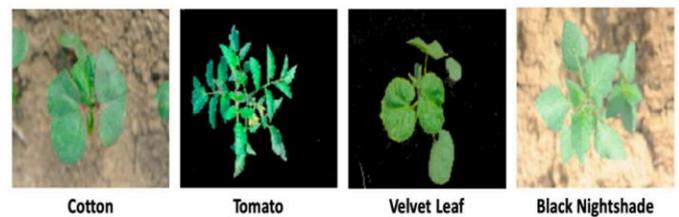


Fig. 2. Sample images of early crop weed detection dataset.

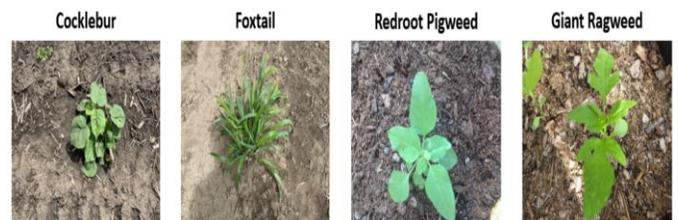


Fig. 3. Sample images of the 4weed dataset.

B. You Only Look Only Once (YOLOv7)

In this research, a computer vision-based recognition and detection technique is provided for object detection. The most recent Yolov7 model was used. Yolov7 is a single-stage object detection technique. You Only Look Once Yolov7's network structure diagram is depicted in Fig. 4 [25]. Overall, the YOLO-V7 technique resizes the input picture to 640x640 before feeding it into the backbone network, producing three layers of feature maps of varying sizes via the head network, and then outputting the prediction result using RepConv [24]. RepConv is utilized to build a planned reparametrized convolution architecture with increased gradient variation for various feature maps [24]. The soft labels generated by the

optimization process are used by the lead head and auxiliary head learning processes, together with the introduction of the auxiliary detecting head. In order to acquire more accurate findings, the soft labels that were produced from it ought to more faithfully represent the distribution and relationship between the source data and the object [26]. Silu activation function, ELAN structures, and MP structures make up the YOLOv7 backbone, by managing gradient pathways and deeper networks, the ELAN structure can effectively learn and converge. Fig. 4 depicts the ELAN and E-ELAN network structures. Down sampling is performed using the MP structure as shown in Fig. 4.

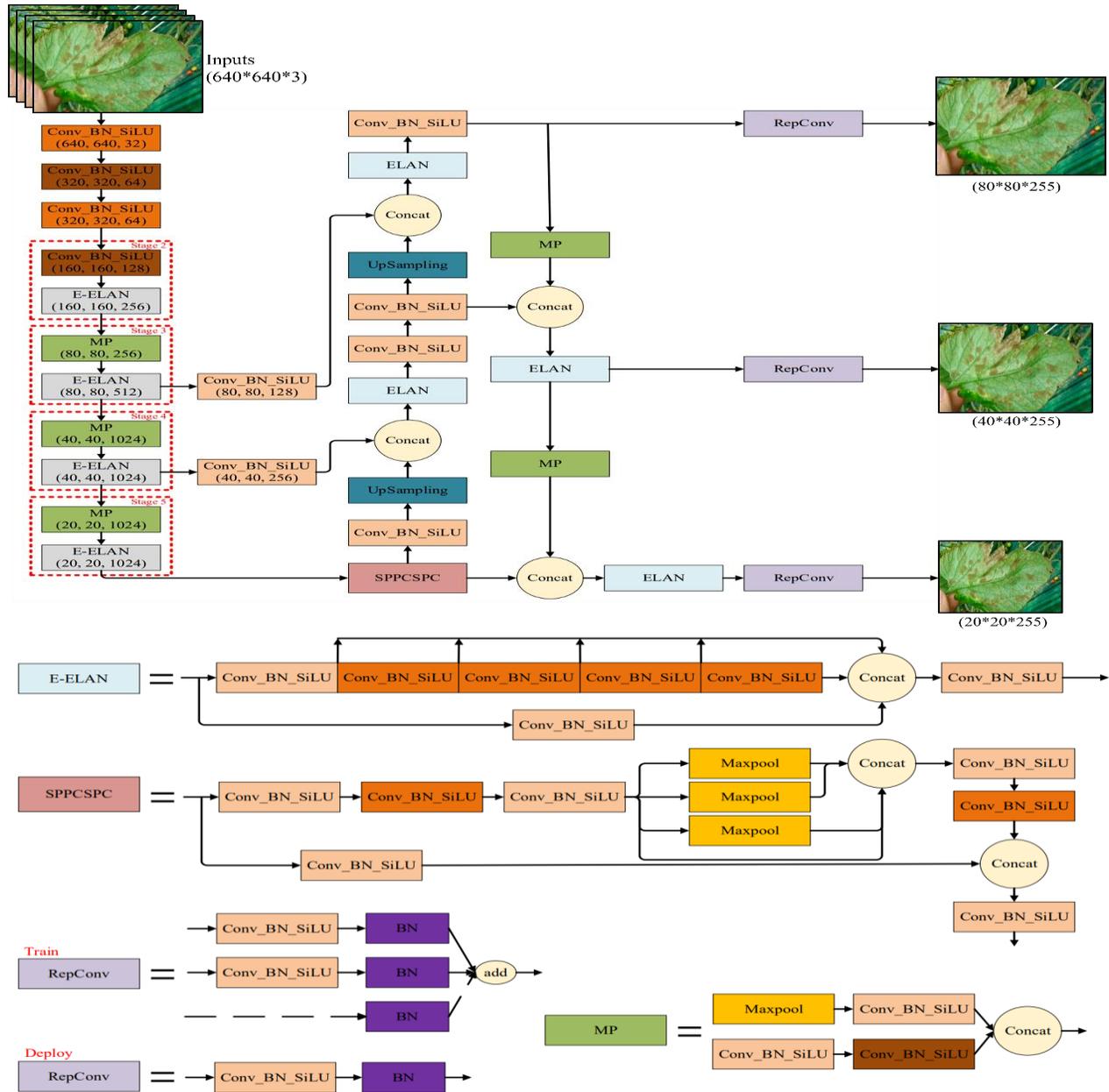


Fig. 4. YOLOv7 architecture.

### III. RESULTS AND DISCUSSION

#### A. Performance Metrics

The most often used metric to evaluate object identification systems are mean average precision (mAP). Comparing the detected box to the corresponding ground truth box allows the mAP to determine its score. The connection between the predicted bounding box coordinates and the actual bounding box is characterized by intersection over union. The projected bounding box coordinates and the truth values should match more closely, according to higher IoU values.

$$IoU = \frac{\text{area of overlap}}{\text{area of union}} \quad (1)$$

The proportion of true positives to all correctly predicted outcomes is referred to as precision. Precision evaluates how accurately a model category a sample as positive.

$$\text{Precision} = \frac{\text{True positives}}{\text{True positives} + \text{False positives}} \quad (2)$$

The proportion of true positives to all the predictions is known as recall. Recall gauges the accuracy with which a model can find positive samples. The most positive samples are discovered when recall is higher.

$$\text{Recall} = \frac{\text{True positives}}{\text{True positive} + \text{False Negatives}} \quad (3)$$

The average area of the precision-recall curve below a given IoU threshold is known as the average precision at IoU ( $AP^{IoU}$ ).  $AP^{IoU}$  is a performance indicator for a certain class or category. To indicate the overall detecting performance, mean average precision at a threshold IoU ( $mAP^{IoU}$ ) is calculated and denoted as follows: =

$$mAP^{IoU} = \frac{1}{N} \sum AP_n^{IoU} \quad (4)$$

$$n \in \{class1, class2, \dots, classN\}$$

One of the often-employed measures for assessing the effectiveness of machine learning algorithms is the F1 score. F1 scores are calculated using the harmonic mean of recall and accuracy. The F1 score value is an indicator of how well categorization systems can anticipate outcomes.

$$F1Score = 2 \times \frac{\text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}} \quad (5)$$

#### B. Performance Evaluation

The model was trained and tested using the cloud-based Google Colab environment, which has access to the NVIDIA Tesla T4 GPU. The YOLOv7 model's training process was started using pre-trained weights derived from the COCO dataset rather than starting from scratch.

We evaluate the effectiveness of the YOLOv7 model developed using the early crop weed dataset and 4Weed dataset. This model trained over 50 epochs. YOLOv7 provides superior mAP than prior trained models, as shown in Table I.

The YOLOv7 model was evaluated using random testing pictures, and the results are displayed in Fig. 5 and Fig. 6. The collected findings suggest that the model can successfully identify agricultural weeds.

TABLE I. COMPARISON OF WEED DETECTION MAP

Model	Dataset	mAP
RetinaNet R101 -FPN[27]	Three weed dataset(Cotton, Carpetweed, Morningglory weed, and Palmer Amaranth weed.)	79.98%
YOLOv5n[27]	Three weed dataset(Cotton, Carpetweed, Morningglory weed, and Palmer Amaranth weed.)	76.58%
Proposed-YOLOv7	Early weed dataset (cotton, velvet weed, tomato, and nightshade weed)	99.6%
Proposed-YOLOv7	4Weed dataset (Giant Ragweed, Foxtail, Cocklebur, and Redroot Pigweed)	78.53%

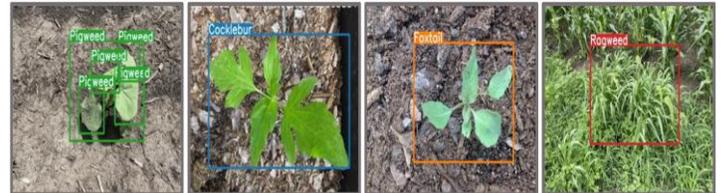


Fig. 5. Early crop weed detection results of YOLOv7.

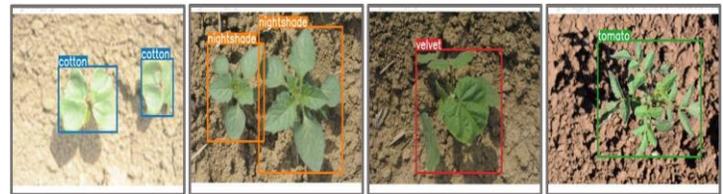


Fig. 6. 4Weed detection results of YOLOv7.

Three different types of losses were produced throughout the YOLOv7 training and validation process: bounding box loss, objectiveness loss, Classification loss, precision, recall,  $mAP@0.5$ , and  $mAP@0.5:0.95$ . Fig. 7 and Fig. 8 show that throughout training, every loss value displayed a decreasing trend, and the model did not exhibit any overfitting. While the validation loss converged near the conclusion of the training, the training loss did so early on. The minimal value in the training and validation loss curves was attained after 50 training epochs with batch size 16.

The Normalized confusion matrix evaluated on test images using YOLOv7 Trained model is plotted as shown in Fig. 9 and Fig. 10.

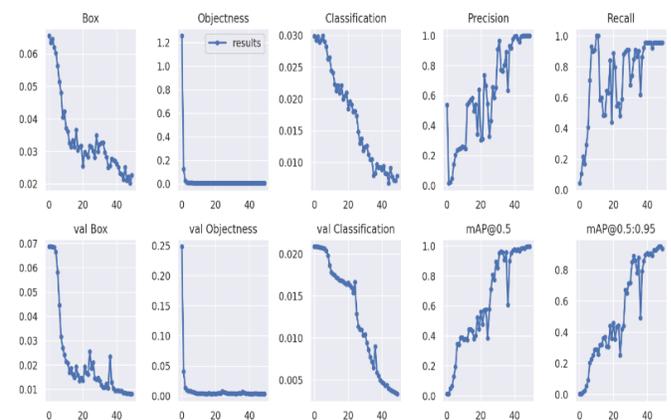


Fig. 7. Performance measures of YOLOv7 on early crop weed dataset.

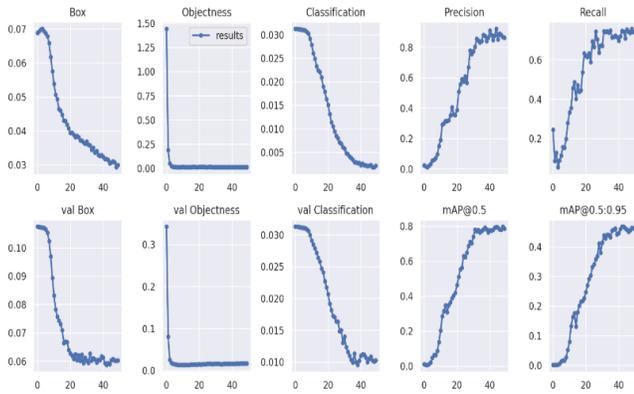


Fig. 8. Performance measures of YOLOv7 on the 4weed dataset.

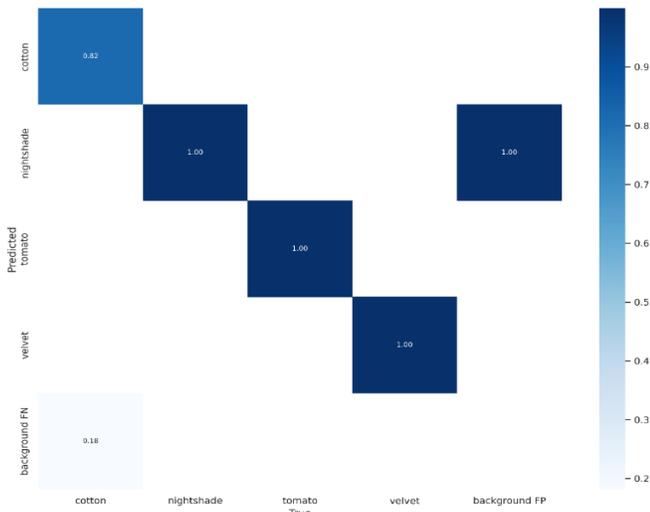


Fig. 9. Confusion matrix of YOLOv7 on early crop weed dataset.

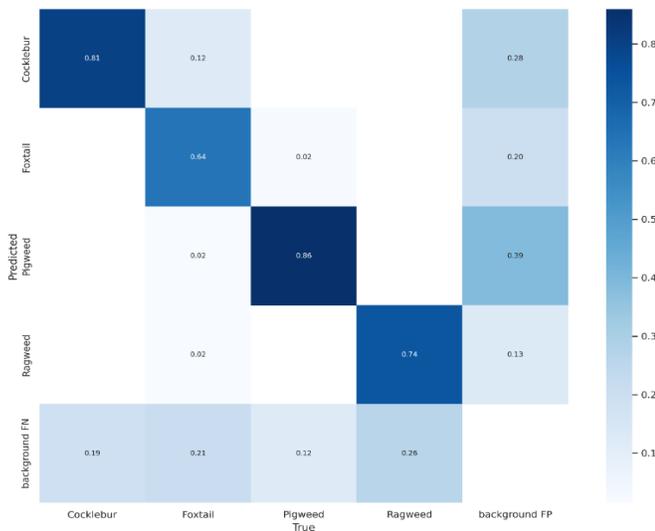


Fig. 10. Confusion matrix of YOLOv7 on 4weed dataset.

The performance of the YOLOv7 model on Early Crop Weed dataset is evaluated using performance metrics mAP, F1score, Precision, Recall, and Precision-Recall. The mAP,

F1score, Precision, and Recall of YOLOv7 on early crop weed dataset after training for 50 epochs are 99.6, 97.6, 99.8, and 95.5. The graphs of these metrics as shown below in Fig. 11 to 14.

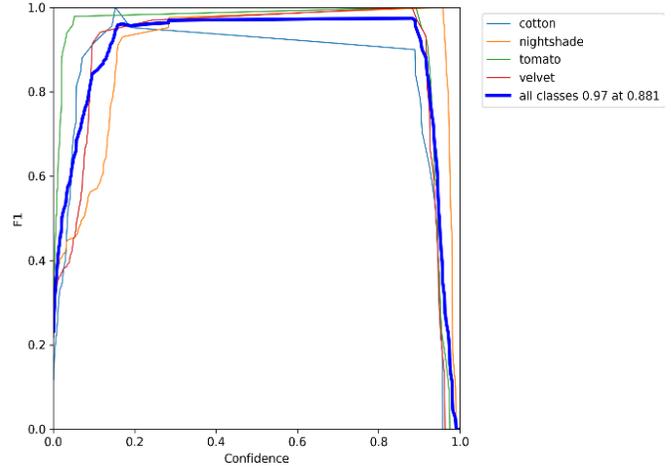


Fig. 11. F1 score of early crop weed dataset.

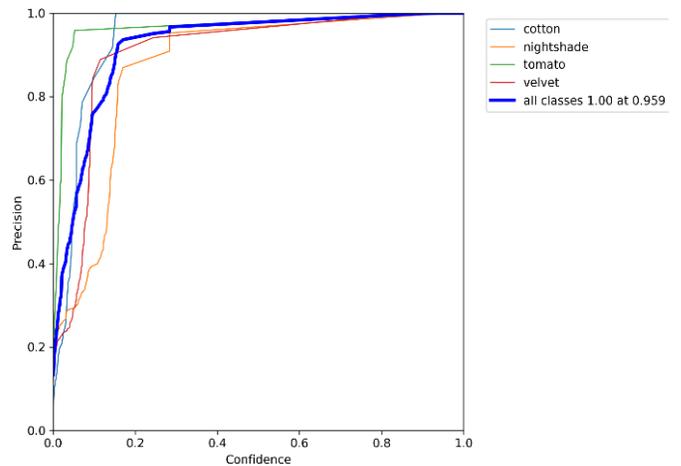


Fig. 12. Precision of early crop weed dataset.

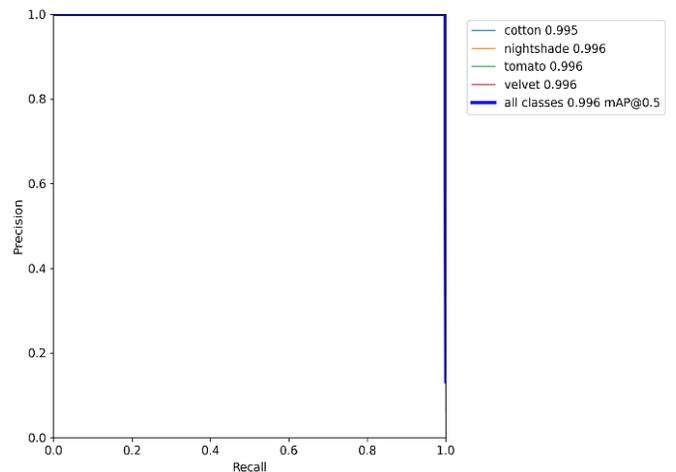


Fig. 13. Recall of early crop weed dataset.

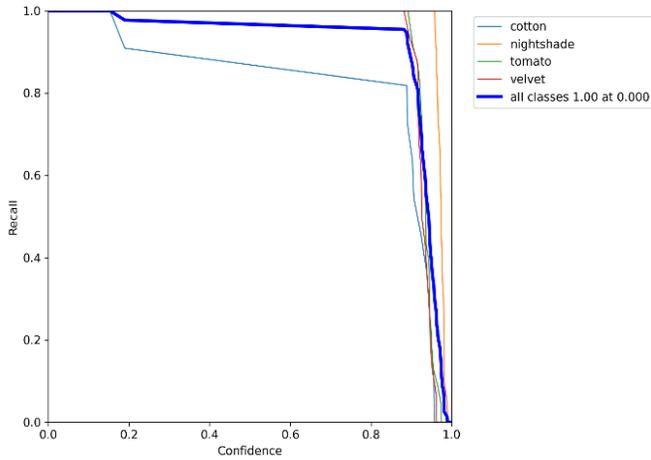


Fig. 14. Precision-recall graph of early crop weed dataset.

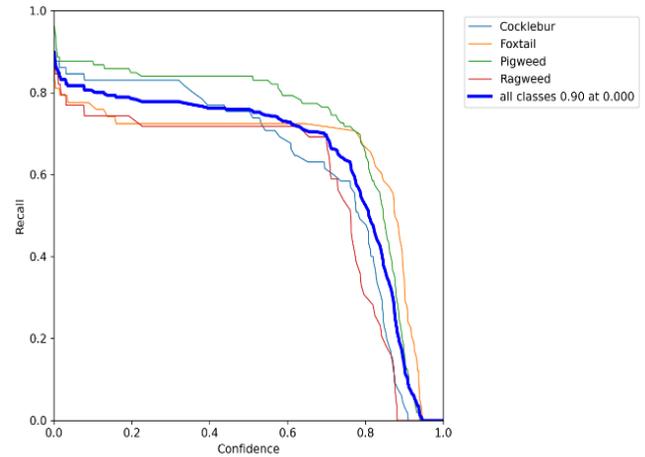


Fig. 17. Recall of 4weed dataset.

The performance of the YOLOv7 model on 4Weed dataset is evaluated using performance metrics mAP, F1score, Precision, Recall, and Precision-Recall. The mAP, F1score, Precision, and Recall of YOLOv7 on the 4weed dataset after training for 50 epochs are 78.53, 79.83, 86.34, 74.24. The graphs of these metrics as shown below in Fig. 15 to 18.

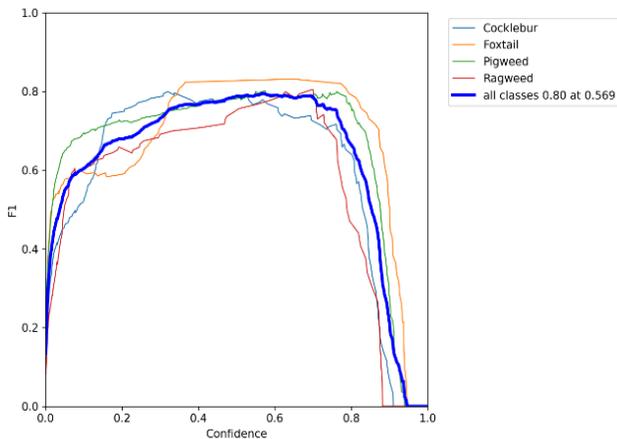


Fig. 15. F1 score of 4weed dataset.

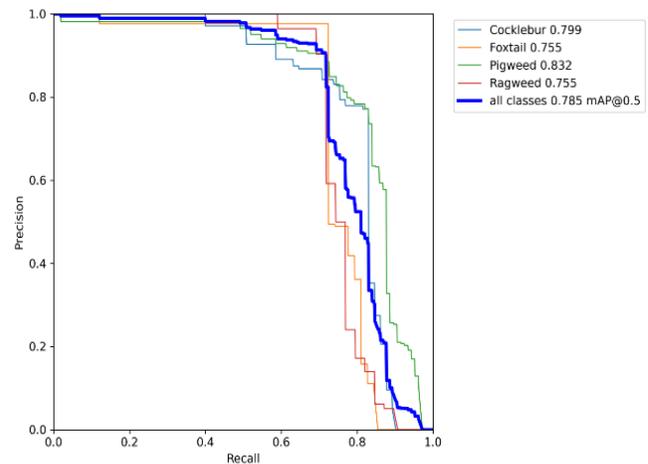


Fig. 18. Precision-recall graph of of 4weed dataset.

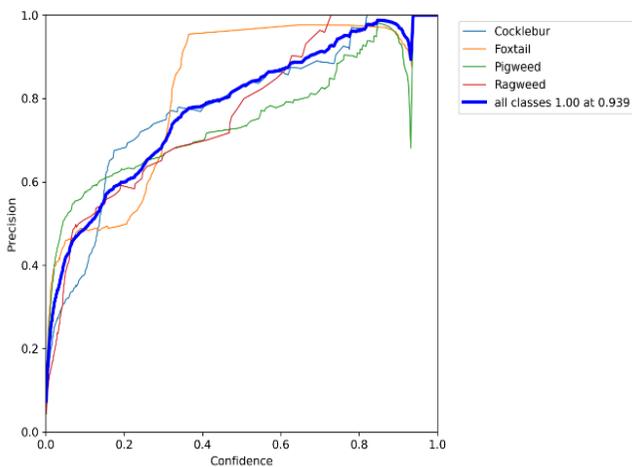


Fig. 16. Precision of 4weed dataset.

The YOLOv7 model have inherent limitations that affect its accuracy for plant weed detection, such as difficulty in detecting small or occluded weeds, or misclassifying non-weed objects as weeds. This study used a limited dataset for training and testing the YOLOv7 model, which could affect the accuracy and generalizability of the results.

#### IV. CONCLUSION

Weeds increase agricultural cultivation costs and lower crop yields. Machine vision plays a significant part in precision agriculture by helping to locate weeds on agricultural land. For the purpose of weed detection using machine vision in this work, we use the early crop weed dataset and the 4weed dataset. On the datasets, the one-stage object detector YOLOv7, which is based on deep learning, was tested for weed detection. The mAP@0.5 detection accuracy for the early crop weed dataset is 99.6 while the mAP@0.5 detection accuracy for the 4weed dataset is 78.53. Because of its quicker inference speed, YOLOv7 has strong promise for real-time applications. By enhancing model training and data augmentation methods, increasing the dataset, and improving the model, further research is still required to increase the accuracy of weed detection. Additionally, field experiments

and demonstrations using trained models deployed on a machine vision system with onboard computer hardware in real-world field settings are required for further model assessment and updating.

#### REFERENCES

- [1] Carballido, J. & Rodríguez-Lizana, A. & Agüera Vega, Juan & Pérez-Ruiz, Manuel. (2013). Field sprayer for inter- and intra-row weed control: Performance and labor savings. *Spanish Journal of Agricultural Research*, 11. 642. 10.5424/sjar/2013113-3812.
- [2] Midtby, Henrik & Mathiassen, Solvejg & Andersson, Kim & Jørgensen, Rasmus. (2011). Performance evaluation of a crop/weed discriminating micro sprayer. *Computers and Electronics in Agriculture*, 77. 35-40. 10.1016/j.compag.2011.03.006.
- [3] Deng, X. & Qi, L. & Ma, X. & Jiang, Y. & Chen, X. & Liu, H. & Chen, W.. (2018). Recognition of weeds at seedling stage in paddy fields using multi-feature fusion and deep belief networks. *Nongye Gongcheng Xuebao/Transactions of the Chinese Society of Agricultural Engineering*, 34. 165-172. 10.11975/j.issn.1002-6819.2018.14.021.
- [4] Ashraf, Taskeen & Khan, Yasir Niaz. (2020). Weed density classification in rice crop using computer vision. *Computers and Electronics in Agriculture*, 175. 105590. 10.1016/j.compag.2020.105590.
- [5] Dadashzadeh, Mojtaba & Abbaspour-Gilandeh, Yousef & Mesri Gundoshmian, T. & Sabzi, Sajad & Hernández Hernández, Jose Luis & Hernández Hernández, Mario & Arribas, Juan. (2020). Weed Classification for Site-Specific Weed Management Using an Automated Stereo Computer-Vision Machine-Learning System in Rice Fields. *Plants*, 9. 559. 10.3390/plants9050559.
- [6] Sharpe, S. M., Schumann, A. W., & Boyd, N. S. (2020). Goosegrass detection in strawberry and tomato using a convolutional neural network. *Scientific Reports*, 10(1), 1-8.
- [7] Veeranampalayam Sivakumar, A. N., Li, J., Scott, S., Psota, E., J. Jhala, A., Luck, J. D., & Shi, Y. (2020). Comparison of object detection and patch-based classification deep learning models on mid-to late-season weed detection in UAV imagery. *Remote Sensing*, 12(13), 2136.
- [8] Jiang, Y., Li, C., Paterson, A. H., & Robertson, J. S. (2019). DeepSeedling: Deep convolutional network and Kalman filter for plant seedling detection and counting in the field. *Plant methods*, 15(1), 1-19.
- [9] Patidar, S., Singh, U., & Sharma, S. K. (2020, July). Weed seedling detection using mask regional convolutional neural network. In 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC) (pp. 311-316). IEEE.
- [10] Lam, O. H. Y., Dogotari, M., Prüm, M., Vithlani, H. N., Roers, C., Melville, B., ... & Becker, R. (2021). An open source workflow for weed mapping in native grassland using unmanned aerial vehicle: Using *Rumex obtusifolius* as a case study. *European Journal of Remote Sensing*, 54(sup1), 71-88.
- [11] Suh, H. K., Ijsselmuiden, J., Hofstee, J. W., & van Henten, E. J. (2018). Transfer learning for the classification of sugar beet and volunteer potato under field conditions. *Biosystems Engineering*, 174, 50-65.
- [12] Huang, H., Deng, J., Lan, Y., Yang, A., Deng, X., Wen, S., ... & Zhang, Y. (2018). Accurate weed mapping and prescription map generation based on fully convolutional networks using UAV imagery. *Sensors*, 18(10), 3299.
- [13] Ma, X., Deng, X., Qi, L., Jiang, Y., Li, H., Wang, Y., & Xing, X. (2019). Fully convolutional network for rice seedling and weed image segmentation at the seedling stage in paddy fields. *PLoS one*, 14(4), e0215676.
- [14] Aggarwal, V., Ahmad, A., Etienne, A., & Saraswat, D. (2022). 4Weed Dataset: Annotated Imagery Weeds Dataset. arXiv preprint arXiv:2204.00080.
- [15] Wu, Z., Chen, Y., Zhao, B., Kang, X., & Ding, Y. (2021). Review of weed detection methods based on computer vision. *Sensors*, 21(11), 3647.
- [16] Lu, Y., & Young, S. (2020). A survey of public datasets for computer vision tasks in precision agriculture. *Computers and Electronics in Agriculture*, 178, 105760.
- [17] Islam, N., Rashid, M. M., Wibowo, S., Xu, C. Y., Morshed, A., Wasimi, S. A., ... & Rahman, S. M. (2021). Early weed detection using image processing and machine learning techniques in an Australian chilli farm. *Agriculture*, 11(5), 387.
- [18] Czymmek, V., Harders, L. O., Knoll, F. J., & Hussmann, S. (2019, May). Vision-based deep learning approach for real-time detection of weeds in organic farming. In 2019 IEEE International Instrumentation and Measurement Technology Conference (I2MTC) (pp. 1-5). IEEE. <https://doi.org/10.1109/i2mtc.2019.8826921>.
- [19] Olsen, A., Konovalov, D. A., Philippa, B., Ridd, P., Wood, J. C., Johns, J., ... & White, R. D. (2019). DeepWeeds: A multiclass weed species image dataset for deep learning. *Scientific reports*, 9(1), 1-12. <https://doi.org/10.1038/s41598-018-38343-3>.
- [20] Hasan, A. M., Sohel, F., Diepeveen, D., Laga, H., & Jones, M. G. (2021). A survey of deep learning techniques for weed detection from images. *Computers and Electronics in Agriculture*, 184, 106067. <https://doi.org/10.1016/j.compag.2021.106067>.
- [21] Espejo-Garcia, B., Mylonas, N., Athanasakos, L., Fountas, S., & Vasilakoglou, I. (2020). Towards weeds identification assistance through transfer learning. *Computers and Electronics in Agriculture*, 171, 105306. <https://doi.org/10.1016/j.compag.2020.105306>.
- [22] Leminen Madsen, S., Mathiassen, S. K., Dyrmann, M., Laursen, M. S., Paz, L. C., & Jørgensen, R. N. (2020). Open plant phenotype database of common weeds in Denmark. *Remote Sensing*, 12(8), 1246. <https://doi.org/10.3390/rs12081246>.
- [23] Sudars, K., Jasko, J., Namatevs, I., Ozola, L., & Badaukis, N. (2020). Dataset of annotated food crops and weed images for robotic computer vision control. *Data in brief*, 31, 105833. <https://doi.org/10.1016/j.dib.2020.105833>.
- [24] Ding, X., Zhang, X., Ma, N., Han, J., Ding, G., & Sun, J. (2021). Reppyg: Making vgg-style convnets great again. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 13733-13742).
- [25] Wang, C. Y., Bochkovskiy, A., & Liao, H. Y. M. (2022). YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. arXiv preprint arXiv:2207.02696.
- [26] Muhammad, M. B., & Yeasin, M. (2020, July). Eigen-cam: Class activation map using principal components. In 2020 International Joint Conference on Neural Networks (IJCNN) (pp. 1-7). IEEE.
- [27] Rahman, A., Lu, Y., & Wang, H. (2023). Performance evaluation of deep learning object detectors for weed detection for cotton. *Smart Agricultural Technology*, 3, 100126.

# Sobel Edge Detection Algorithm with Adaptive Threshold based on Improved Genetic Algorithm for Image Processing

Weibin Kong, Jianzhao Chen, Yubin Song, Zhongqing Fang, Xiaofang Yang, Hongyan Zhang  
School of Information Engineering-Research Center of Photoelectric and Information Technology,  
Yancheng Institute of Technology, Jiangsu Yancheng, China

**Abstract**—In this paper, a novel adaptive threshold Sobel edge detection algorithm based on the improved genetic algorithm is proposed to detect edges. Because of the influence of external factors in actual detection process, the result of detection is often not accurate enough when the configured threshold of the target image is far away from the real threshold. Different thresholds of images are calculated by improved genetic algorithm for different images. The calculated threshold is used in edge detection. The experimental results show that the image processed by the improved algorithm has stronger edge continuity. It is shown that proposed algorithm has a better detection effect and applicability than the traditional Sobel algorithm.

**Keywords**—Genetic algorithm; Sobel operator; edge detection; adaptive threshold

## I. INTRODUCTION

As one of the most important parts of image processing, edge detection is of great significance to image high-order feature extraction, target recognition, image segmentation and many other fields. The image edge refers to the area where the grey value of adjacent pixels changes dramatically [1]. Various edge detection methods have been proposed. According to different detection principles, it can be divided into first-order differential operator, second-order differential operator and modern edge detection algorithm. The first-order differential operators include Sobel operator, Robert's operator, Prewitt operator, etc. The second-order differential operators include Laplace operator, Canny operator, LOG operator, etc. Modern edge detection algorithms include wavelet transform, neural network algorithm, etc. [2-3].

For different edge detection algorithms, many optimization strategies have been proposed to obtain better detection effects. Image edge detection is greatly affected by noise. In the process of Sobel edge detection, the mean filtering method of median filtering is commonly used to remove noise. However, this method has a general effect of removing salt and pepper noise. To better remove noise, soft threshold wavelet denoising is applied to Sobel operator edge detection[4]. In order to meet the real-time requirements of PC ports, an eight-direction adaptive threshold Sobel operator edge detection algorithm is proposed, which's mapping is realized on FPGA [5]. The adaptive threshold edge detection algorithm based on fuzzy divergence uses adaptive threshold to detect the target image [6].

With the rise of artificial intelligence, the application of artificial intelligence algorithms in the field of image processing has become a hot topic. In aerospace, remote sensing, medical and other high-end industries, the accuracy of image processing is required to be high. In view of this situation, artificial intelligence algorithm has great advantages. In addition, deep learning algorithm and neural network algorithm are applied in edge detection [7-9]. In the processing of complex images, artificial intelligence algorithm has advantages. However, the artificial intelligence algorithm has the disadvantages of large amount of calculation, long time for edge detection and the need to train the algorithm in advance when detecting the edge of specific image. In practical applications, the edge detection algorithm is required to have the characteristics of fast response, low hardware requirements, wide range of detection targets and so on.

The research of traditional image edge detection is still widely concerned. For the limited edge extraction of Sobel operator in horizontal and vertical directions, the quantum of non-maximum suppression and double threshold technology are adopted to improve it [10]. For the noise problem in Sobel operator edge detection, median filter algorithm, mean filter algorithm, weighted kernel norm minimization image denoising algorithm and traditional Sobel algorithm are usually used to achieve better detection effect [11]. However, the Sobel edge detection algorithm still needs to manually set the threshold [12].

To solve this problem, an adaptive threshold edge detection algorithm based on improved genetic algorithm is proposed. According to different target images, different thresholds are calculated and used for edge detection to avoid the problem of low accuracy of edge detection caused by improper manual threshold setting.

## II. PRINCIPLE OF SOBEL EDGE DETECTION

As a common edge detection algorithm, Sobel operator combines Gaussian smoothing and differentiation to calculate the gradient value of image brightness function. The main method is to make the weighted difference between the gray values of the upper, lower, left and right fields of the target pixels, and then smooth the image. The greater the weight close to the target pixel, the greater the impact on the target during convolution [13-14]. The steps of the Sobel operator to determine the image edge are as follows:

Step 1: Horizontal operator  $Sobel_x$  and vertical operator  $Sobel_y$  are used to convolute the target image. The horizontal and vertical operators of Sobel algorithm are described as follows:

$$Sobel_x = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix} Sobel_y = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix} \quad (1)$$

Step 2: The convolution value of Sobel operator in  $x$  and  $y$  directions is calculated by:

$$\begin{aligned} G_x(x, y) &= [F(x+1, y-1) + 2F(x+1, y) + F(x+1, y+1) \\ &\quad - F(x-1, y-1) - 2F(x-1, y) - F(x-1, y+1)] \\ G_y(x, y) &= [F(x-1, y+1) + 2F(x, y+1) + F(x+1, y+1) \\ &\quad - F(x-1, y-1) - 2F(x, y-1) - F(x+1, y-1)] \end{aligned} \quad (2)$$

where  $G_x$  and  $G_y$  are the convolution value in  $x$  and  $y$  direction, respectively.  $F(x+n, y-m)$  refers to the gray value of the point. Then the gradient vector is calculated by:

$$G = \sqrt{G_x^2 + G_y^2} \quad (3)$$

Step 3: By comparing with the preset threshold, it is determined that the point is an image edge when the value of the point is greater than the threshold, otherwise is not.

The classic Sobel edge detection operator uses the gray weighting algorithm of the top, bottom, left and right of the target pixel to determine whether it is an edge pixel. This method can not only extract the edge of the target image but also smooth the noise. However, the effect is not very well when the edge extraction is fine or the image threshold is not set properly.

### III. IMPROVED SOBEL EDGE DETECTION ALGORITHM

If the threshold setting in the traditional Sobel operator edge detection algorithm is unreasonable, the accuracy of edge detection is not high. However, the threshold setting is manual, and different images need to be set according to experience. In this way, a large number of pictures cannot be processed automatically. In this paper, adaptive genetic algorithm and traditional Sobel edge detection algorithm are combined. Different thresholds of images are calculated by genetic algorithm for different images. The calculated threshold is used in edge detection. The process of the improved Sobel edge detection algorithm is shown in Fig. 1.

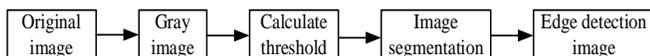


Fig. 1. Flow chart of improved Sobel edge detection algorithm.

#### A. Improved Sobel Edge Detection Algorithm

The genetic algorithm is a computational model of the biological evolution process that simulates for the natural

selection and genetic mechanism of Darwin's biological evolution theory. It is the method to search the optimal solution by simulating the natural evolution process. Genetic algorithm was proposed by Professor Holland firstly. It is a general solution to solve the search algorithm, especially suitable for the calculation of the optimal solution. The genetic algorithm has been widely used in many fields such as production scheduling, automatic control, image processing, machine learning and so on [15-17].

The basic operation of the genetic algorithm includes three basic operators: selection, crossover, and mutation. The genetic operation of an individual population is carried out under random disturbance, so the migration process of the optimal solution is also random. It should be emphasized that this random migration is different from the traditional random search, and the search of genetic operation is efficient and directional. The specific manifestations of genetic operation are as follows:

1) *Selection*: The operation of choosing high-quality individuals in the population and eliminating low-quality individuals is called selection. The purpose of the selection operation is to transfer the genes of high-quality individuals to the next generation. The criterion of selection is based on an appropriate individual fitness assessment. At present, the commonly used fitness evaluation methods include the roulette selection method, fitness proportion method, random traversal sampling method, local selection method and others. And roulette selection method is used most widely.

2) *Crossover*: In the process of biological evolution, genetic recombination plays a key role. Similarly, in the genetic algorithm, crossover operator operation is the core of the whole genetic algorithm. Cross operation means recombining part structure of two-parent individuals to obtain a new individual through cross replacement and achieving the purpose of recombining the desired gene. The crossing modes include single-point crossing and multi-point crossing and uniform crossing. The most common is single-point crossing.

3) *Mutation*: The operation of the mutation operator in the genetic algorithm is to change a certain gene of an individual and to achieve the purpose of changing individual genes. The mutation operation not only increases the diversity of the population and prevents premature convergence, but also enhances the local search ability of the genetic algorithm and accelerates the convergence. Common mutation methods include real-valued mutation and binary mutation.

The above operations are the basic operations of the genetic algorithm. The effect of the genetic algorithm is largely related to the population size, iteration times, crossover probability and mutation probability set by the three genetic operators. In this term, this paper adopts an adaptive genetic algorithm compared with the traditional genetic algorithm. By calculating individual fitness, the algorithm adopts different crossover and mutation probabilities for individuals with different fitness, to increase the individual diversity of the population and achieve the purpose of rapid global convergence. The specific improvement details are as follows.

1) *Setting different crossover probabilities according to the individual fitness of the population:* For individuals with the highest fitness do not select cross, and their genes are directly transmitted to the new offspring. For the individuals with the lowest fitness, selecting complete crossover and covering their original genes, then the fitness of their offspring is greatly improved. For other individuals, the crossover probability is determined according to the specific situation. This improved crossover operator not only ensures the individual diversity of the population, but also makes the individual fitness converge rapidly in the direction of the optimal solution. The mathematical expression of the improved cross-operation principle is defined as follows:

$$P_c = \begin{cases} k_2, & f_c = f_{\min} \\ k_1 \frac{f_{\max} - f_c}{f_{\max} - f_{\min}}, & f_c \neq f_{\max}, f_{\min} \\ k_3, & f_c = f_{\max} \end{cases} \quad (4)$$

where  $P_c$  is the probability of crossover, and  $f_c$  is the individual with greater fitness in the two individuals of the parent generation.  $f_{\max}$  and  $f_{\min}$  are the maximum and minimum fitness of the individual in the population.  $k_1, k_2, k_3$  are constants between 0 and 1, and  $k_2 > k_3$ .

2) *Adaptive mutation probability:* Mutation operation is an indispensable link in genetic algorithm, which mainly ensures the diversity of population genes. Through the combined action of crossover and mutation, the population will rapidly converge towards the optimal solution. In the mutation operation, the mutation probability is used to represent the intensity of mutation. Usually, a smaller value is required to prevent the mutation operation from misoperating the genes of excellent individuals in the population. In order to prevent the premature convergence of the population caused by the misoperation of the mutation operation on the excellent individuals, the adaptive improvement is made in this paper. The mathematical principle is shown:

$$P_m = \begin{cases} k_5, & f_c = f_{\min} \\ k_4 \frac{f_{\max} - f_c}{f_{\max} - f_{\min}}, & f_c \neq f_{\max}, f_{\min} \\ k_6, & f_c = f_{\max} \end{cases} \quad (5)$$

where  $P_m$  is the mutation probability,  $f_c$  is the one with greater adaptability among parents, and  $f_{\max}, f_{\min}$  are the maximum fitness and minimum fitness of the population.  $f_4, f_5, f_6$  are constants between 0 and 1, and  $k_5 > k_6$ . Through the adaptive improvement of crossover operation and mutation operation. The flow of the adaptive genetic algorithm is shown in Fig. 2.

### B. The Principle of Calculating the Optimal Threshold by Genetic Algorithm

In the process of using genetic algorithm to calculate the image threshold, the maximum inter class variance of the image is used as the best fitness function to calculate the threshold.

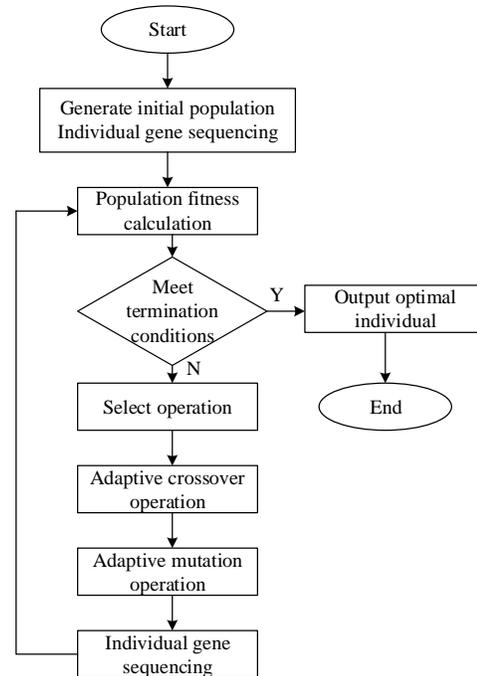


Fig. 2. Flow chart of improved genetic algorithm.

Let  $(x, y)$  be the coordinate of a pixel on the image, and the value range of the gray level of the image is  $G = \{0, 1, 2, \dots, L-1\}$ , where 0 represents the darkest pixel and  $L-1$  represents the brightest pixel. The gray level of the point whose coordinates are  $(x, y)$  is  $f(x, y)$ . Let  $t \in G$  the segmentation threshold,  $B = \{C_0, C_1\}$  be a binary gray level, and  $C_0, C_1 \in G_0$ , then the expression result of the function  $f(x, y)$  on the threshold  $t$  is:

$$f_t(x, y) = \begin{cases} C_0 & f(x, y) \leq t \\ C_1 & f(x, y) > t \end{cases} \quad (6)$$

If the number of pixels with gray  $i$  is  $m_i$ , the total number of image pixels is:

$$M = \sum_{i=1}^{L-1} m_i \quad (7)$$

The probability of occurrence of gray level  $i$  is

$$P_i = \frac{m_i}{M} \quad (8)$$

The traditional maximum inter class variance method is used for image segmentation. Let  $f(x, y)$  be the image to be segmented. The gray scale range of the image is  $\{0, 1, 2, \dots, L-1\}$ . The threshold  $t$  divides the pixels in the image into two categories:  $C_0 = \{0, 1, \dots, t\}$ ,  $C_1 = \{t, t+1, \dots, L-1\}$ .  $C_0$  and  $C_1$  represent the target and background respectively.

Normalize the histogram of the image to obtain the probability distribution of the gray level as follow:

$$P_i = n_i / N, P_i \geq 0, \sum_{i=0}^{L-1} P_i = 1 \quad (9)$$

where  $n$  is the number of pixels with gray scale  $i$ ,  $N$  is the number of all pixels of the image,  $P_i$  is the probability of occurrence of gray level  $i$ . The probability of occurrence of  $C_0$  and  $C_1$  is:

$$\begin{aligned} \omega_0 &= \sum_{i=0}^t n_i / N = \sum_{i=0}^t P_i \\ \omega_1 &= \sum_{i=t+1}^{L-1} n_i / N = \sum_{i=t+1}^{L-1} P_i = 1 - \omega_0 \end{aligned} \quad (10)$$

The mean values of  $C_0$  and  $C_1$  are respectively.

$$\begin{aligned} \mu_0 &= \sum_{i=0}^t n_i * i / \sum_{i=0}^t n_i = \sum_{i=0}^t P_i * i / \omega_0 \\ \mu_1 &= \sum_{i=t+1}^{L-1} n_i * i / \sum_{i=t+1}^{L-1} n_i = \sum_{i=t+1}^{L-1} P_i * i / \omega_1 \end{aligned} \quad (11)$$

Let  $\mu$  be the mean value of the whole image,  $\mu = \sum_{i=0}^{L-1} P_i * i$ .

When the threshold is  $t$ , the gray value is  $\mu_t = \sum_{i=0}^t P_i * i$ . The average value of sampled grayscale is  $\mu = \mu_0 \omega_0 + \mu_1 \omega_1$ . The variance between the two classes is:

$$\sigma^2 = \omega_0 (\mu_0 - \mu)^2 + \omega_1 (\mu_1 - \mu)^2 = \omega_0 \omega_1 (\mu_0 - \mu_1)^2 \quad (12)$$

when  $\sigma^2$  is maximum,  $t$  is the optimal threshold.

### C. Adaptive Threshold Sobel Edge Detection Algorithm

The traditional Sobel needs to manually set the image threshold. When the manually set threshold is close to the real image threshold, the effect of image edge detection is preferably. However, in practice, image acquisition is greatly affected by external factors such as illumination, which makes the threshold values of different images vary greatly. The effect of traditional Sobel operator in image edge detection with different thresholds is not ideal.

To solve this problem, this paper combines the genetic algorithm with the traditional Sobel edge detection algorithm and proposes an adaptive threshold Sobel edge detection algorithm. Setting different thresholds for different images avoids the influence of external factors on image segmentation and has a better effect. The implementation steps of the algorithm are as follows:

- 1) Import the target image, determine the parameter set according to the actual problem, encode the parameter set, set the initial parameters, etc.
- 2) The population individuals are selected, crossed, and mutated adaptively and iterated to the optimal individuals, which means calculating the threshold of the image.
- 3) Pass the threshold in step 2) to the classical Sobel edge detection algorithm to avoid manually setting the threshold.
- 4) Compare the threshold with the convolution value calculated by the convolution operator to judge the image edge.
- 5) Complete edge detection and output the target image.

The above is the whole process of the adaptive threshold Sobel edge detection algorithm.

## IV. EXPERIMENTAL RESULTS AND ANALYSIS

In this section, we present the results by the proposed algorithm and the traditional Sobel algorithm.

### A. Initial Parameter Setting

Take the person image shown in Fig. 6(a) as the test image to test the edge detection process of the algorithm. The initial parameter settings are shown in Table I.

The three groups of initial parameters in Table I are tested respectively to obtain the best fitness curve and the best threshold curve under the three groups of parameters. The best fitness curve and the best threshold curve of the three groups of parameters are shown in Fig. 3, Fig. 4, and Fig. 5.

TABLE I. INITIAL PARAMETERS OF CHARACTER IMAGE EDGE DETECTION

Initial parameters	Population quantity	Number of iterations	Crossover probability	Variation probability
Group 1	10	100	0.8	0.5
Group 2	5	50	0.6	0.4
Group 3	20	200	0.9	0.7

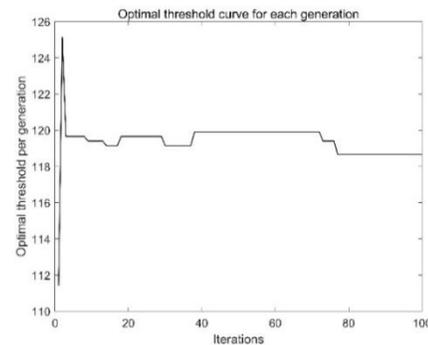


Fig. 3. Group 1's optimal threshold.

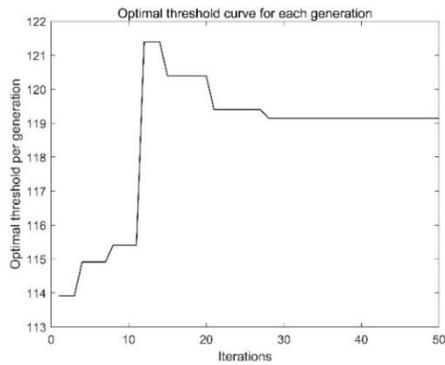


Fig. 4. Group 2's optimal threshold.

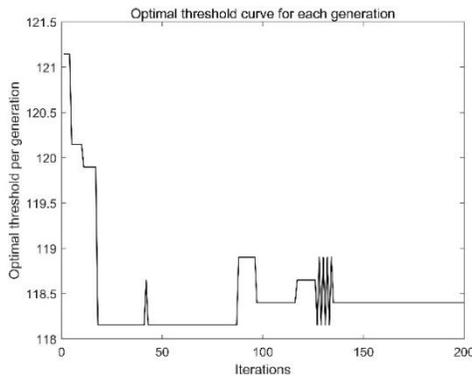


Fig. 5. Group 3's optimal threshold.

In group 1, the optimal threshold is 118. In group 2, the optimal threshold is 120. In group 3, the optimal threshold is 118. In the group 2 of data, the optimal solution will converge in advance due to the small number of populations and iterations. The calculation of image threshold is inaccurate. Although the image threshold calculation of the group 3 of data is accurate, the amount of calculation increases and resources are wasted due to the excessive parameter setting. Comprehensive comparison shows that the group 1 of data not only ensures the accuracy of calculation, but also has a moderate amount of calculation. Therefore, in the following image processing, the group 1 of parameters are used as the original parameters to calculate the threshold of the image.

### B. Edge Detection Process

According to the initial parameters of group 1, the threshold value of the image in Fig. 6(a) is calculated, which is 118. Before edge detection of the target image, the image needs to be processed. First, the grayscale image shown in Fig. 6(b) is obtained by grayscale processing of Fig. 6(a). In order to make the experimental process closer to the real detection process, random salt and pepper noise is added to the gray image to simulate the random noise in the process of image acquisition. The noise diagram shown in Fig. 6(c) is obtained. After the above preparations are completed, the threshold value detected in Section 4A is transferred to Sobel operator to detect the edge of the target image. Finally, the edge detection effect shown in Fig. 6(d) is obtained.

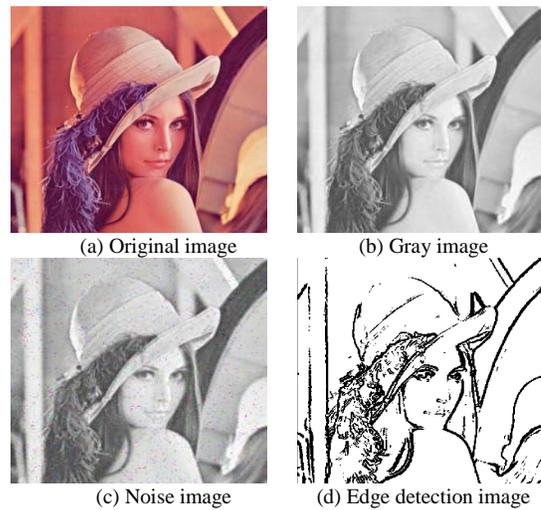


Fig. 6. Edge detection process.

As shown in Fig. 6, the algorithm in this paper has a good effect on the edge detection of the target in the noisy environment. The contour of edge detection is clear. The image details are complete. The image information is completely saved, which can meet the further processing of the target image. At the same time, the algorithm overcomes the problem that the edge detection effect is not ideal due to the large gap between the image threshold and the set threshold. The threshold of the target image is calculated by genetic algorithm. The calculated threshold value is then transferred to the Sobel operator to replace the manually set threshold value. Finally, the purpose of adaptive threshold edge detection is achieved.

### C. Analysis of Experimental Results

In order to show the edge detection effect of the algorithm under different images, this paper uses three examples to compare the algorithm with the traditional Sobel algorithm with different thresholds. The traditional Sobel operator setting threshold and adaptive threshold of each image are shown in Table II. The comparison diagram of different thresholds of the three examples is shown in Fig. 7, Fig. 8, and Fig. 9.

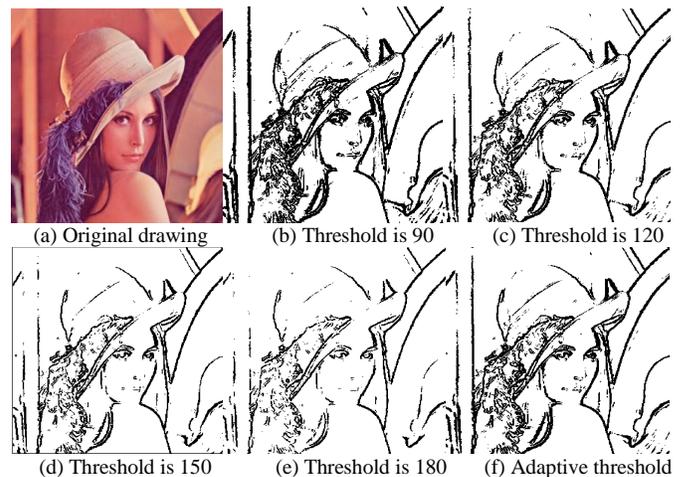


Fig. 7. The example 1 of edge detection of human image under different thresholds.

TABLE II. THRESHOLD SETTING OF DIFFERENT IMAGES

Image name	Example 1	Example 2	Example 3
Threshold 1	90	90	90
Threshold 2	120	120	120
Threshold 3	150	150	150
Threshold 4	180	180	180
Adaptive threshold	118	143	88

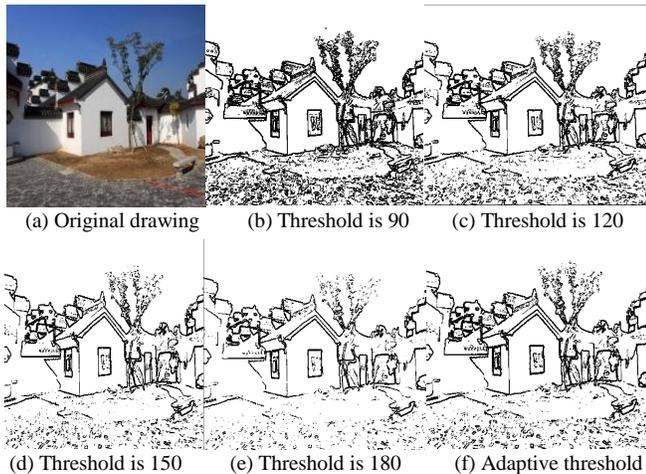


Fig. 8. The example 2 of edge detection of building image under different thresholds.

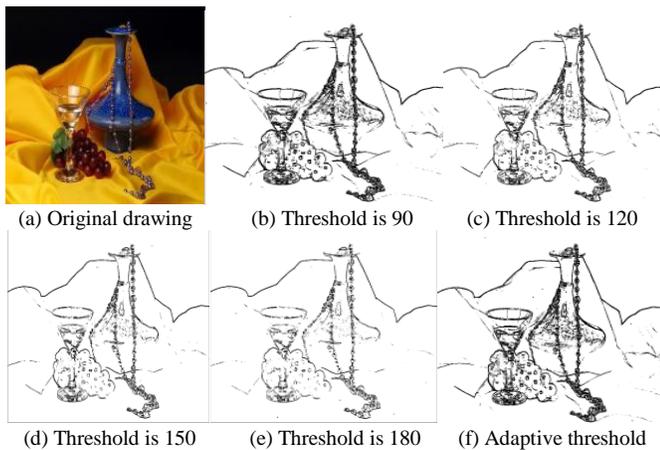


Fig. 9. The example 3 of edge detection of still life image under different thresholds.

## V. CONCLUSIONS

In this paper, Sobel edge detection algorithm based on the genetic algorithm is proposed. Compared with the traditional Sobel algorithm, it does not need to set the threshold manually, which avoids the poor detection effect caused by different image thresholds. This algorithm has great improvement. The image edge is located accurately. The image details are well

preserved. It can adapt to images with different thresholds and has great practical application value.

## ACKNOWLEDGMENT

This work was supported by the Graduate Innovation Project (SJCX22-XZ033), Yancheng Institute of Technology Teaching Reform Research Project (JYKT2022A028).

## REFERENCES

- [1] D. R. Waghule, R. S. Ochawar, "Overview on edge detection methods," in: Proceedings of the 2014 International Conference on Electronic Systems, Signal Processing and Computing Technologies. 2014,151-155.
- [2] H. X. Zhang, C. Wan, X. Liu, "Research progress of image edge detection algorithm," Comput. En. Appt. 2018,54(14):11-18.
- [3] D. Lv, S. H. Pan, "Improved Canny edge detection algorithm based on deep learning Scientific," J. Intel. Syst. Research. 2021,3(2):114-118.
- [4] P. Chen, Y. Wang, J. Xing, "Research on design and verification of Sobel image edge detection based on high level synthesis," in: Proceedings of the 2019 6th International Conference on Dependable Systems and Their Applications. 2020,491-492.
- [5] X. X. Zou, Y. H. Zhang, S. Y. Zhang, J. Zhang, "FPGA implementation of edge detection for Sobel operator in eight directions," in: Proceedings of the 2018 IEEE Asia Pacific Conference on Circuits and Systems. 2018, 520-523.
- [6] M. Versaci, F. C. Morabito, "Image edge detection: a new approach based on fuzzy entropy and fuzzy divergence," Int. J. of Fuzzy Systems. 2021,23(4):918-936.
- [7] Y. B. Yu, C. Y. Yang, Q. X. Deng, T. Nyima, S. Y. Liang, C. Zhou, "Memristive network-based genetic algorithm and its application to image edge detection," J. Syst. Eng. Electron. 2021,32(5),1062-1070.
- [8] H. Zhao, B. Wu, Y. Guo, "SSWS: An edge detection algorithm with strong semantics and high detectability for spacecraft," Optik. 2021, 247, 168037.
- [9] V. Maksimovic, M. Petrovic, D. Savic, "New approach for estimating edge detection threshold and application of adaptive detector depending on image complexity," Optik. 2021,238:166476.
- [10] R. Chetia, S. M. B. Boruah, P. P. Sahu, "Quantum image edge detection using improved Sobel mask based on NEQR. Quantum," Inf. Process. 2021,20(1),1-25.
- [11] R. Tian, G. Sun, X. Liu, "Sobel edge detection based on weighted nuclear norm minimization image denoising," Elec. 2021,10(6),655.
- [12] S. N. Ning, M. Zhu, H. H. Sun, "FPGA implementation of an improved Sobel adaptive edge detection," Chinese. J. Liq. Cryst. Displays. 2014, 29(03): 395-402.
- [13] N. Nausheen, A. Seal, "A FPGA based implementation of Sobel edge detection," Microprocess. Microsy. 2018,56,84-91.
- [14] P. Kanchanatripop, D. Zhang, "Adaptive Image Edge Extraction Based on Discrete Algorithm and Classical Canny Operator," Symmetry. 2020, 12(11):1749.
- [15] B. Xu, H. Q. Zhu, X. Wang, "Decoupling control of outer rotor coreless bearing less permanent magnet synchronous motor based on least squares support vector machine generalized inverse optimized by improved genetic algorithm," IEEE. T. Ind. Electron. 2022, 69(12), 12182-12190.
- [16] M. M. Li, C. S. Zou, "Threshold image segmentation method based on improved genetic algorithm," Software. Eng. 2022, 25(01),37-40.
- [17] S. Mirjalili, Evolutionary Algorithms and Neural Networks: Theory and Application. Switzerland: Springer, 2019.

# Compiler Optimization Prediction with New Self-Improved Optimization Model

Chaitali Shewale<sup>1</sup>, Sagar B. Shinde<sup>2</sup>, Yogesh B. Gurav<sup>3</sup>, Rupesh J. Partil<sup>4</sup>, Sandeep U. Kadam<sup>5</sup>

Vishwakarma Institute of Information Technology, Pune<sup>1</sup>

Dr. D. Y. Patil Institute of Technology, Pimpri, Pune<sup>2</sup>

Navsahyadri College of Engineering, Pune<sup>3,4</sup>

Anantrao Pawar College of Engineering & Research, Pune<sup>5</sup>

**Abstract**—Users may now choose from a vast range of compiler optimizations. These optimizations interact in a variety of sophisticated ways with one another and with the source code. The order in which optimization steps are applied can have a considerable influence on the performance obtained. As a result, we created a revolutionary compiler optimization prediction model. Our model comprises three operational phases: model training, feature extraction, as well as model exploitation. The model training step includes initialization as well as the formation of candidate sample sets. The inputs were then sent to the feature extraction phase, which retrieved static, dynamic, and improved entropy features. These extracted features were then optimized by the feature exploitation phase, which employs an improved hunger games search algorithm to choose the best features. In this work, we used a Convolutional Neural Network to predict compiler optimization based on these selected characteristics, and the findings show that our innovative compiler optimization model surpasses previous approaches.

**Keywords**—Compiler optimization prediction; feature extraction; feature exploitation; improved hunger games search algorithm; convolutional neural network

## I. INTRODUCTION

As per Moore's Law, the density of transistors doubles every 2 years. Compilers, on the other hand, progress at a pace of a couple percent of year. Compilers were vital tools for connecting written software to destination hardware. In the field of compilers, there's several unresolved research issues [1]. Compilers play a crucial role in software development. Its core objective is to boost software productivity [2].

Compilers were liable for two tasks: translation as well as optimization. They must first effectively convert programmes into binary. Secondly, they must discover the most cost-effective translation. There are numerous valid translations, each of which performs distinctively. The great majority of studies and technological activities are centered on this second performance objective, which has been referred to as optimization. The objective was mislabeled because, until recently, most people rejected obtaining an ideal translation as a difficult and impractical task [3]. Compilers are now being improved so that every code block in a programmed may be transformed into an efficient application [4]. Traditional compiler optimization seems to be a difficult process with no assurances of producing the most effective and quickest target code [5]. A compiler enables a multitude of code optimizations that could be activated or disabled via a compilation flag in

order to enhance the throughput of compiled applications. Nevertheless, because the influence of compiler optimizations largely dependent on programme features (e.g., programme structures), the identical optimizations may not surely result in the identical runtime speed boost when implemented to various programmes [6].

Furthermore, there's an infinite range of flag combos owing to the enormous count of optimization flags. Users may find it difficult to comprehend all of the flags including their combos, and to correctly decide which flags should be activated or disabled in attempt for built programmes to attain the desired runtime performance [7]. Compilers for machine learning (ML) tackle a lot of optimization issues in order to convert an ML programmed, which is often expressed as a tensor computational graph, into an efficiently executable for a hardware destination [8]. Prior efforts [9] – [14] have permitted optimizations that are implemented at the very same point in the compilation pipeline, notably the loop conversion phase. However, since compiler modifications are arranged as passes to minimize complication and also have rigorous ordering limitations, this is unfeasible in production compilers.

With a compiler's optimization capabilities influencing so many parts of product development, understanding and evaluating a compiler's optimization technology is more critical than ever. In this work, an improved optimization prediction model was created, which not only decreases computational time but also enables the compilers with faster convergence, more stable balance, and high-quality outcomes by selecting appropriate optimization. Our work made the following contributions:

- Several high-level characteristics may arise from the coefficients as a result of improved entropy extraction, which boosts the compiler optimization prediction.
- An Improved Hunger Game Search optimization was proposed to provide a very competitive performance to the compiler with less computational time.

The following is the flow of this article: Section II covers some previous relevant research, Section III gives a brief presentation of our proposed compiler optimization prediction model, Section IV gives the outcomes of our work, and Section V contains the conclusion, while the following section includes the references for this work.

## II. RELATED WORKS AND REVIEW

Some of the researches presented by various researchers on compiler optimization were reviewed here.

Hui et al. [15] presented the ALIC iterative compiler optimization parameters estimation model, which has minimal overheads. Firstly, the target programmes were defined using static-dynamic characteristics format depending on feature significance, as well as an early optimization prediction model was built using the classifier. Subsequently, for every sample, a dynamic amount of sample observation methodology was being used. The most beneficial test from the collection of candidate samples typically the chosen and labeled with each mark increase the count of sample data. The optimization prognosis system is then built by using intermediate prediction network, which actively learns candidate samples.

Tiago et al. [16] suggested a new exploration approach to determine a compiler optimization strategy. This hybrid methodology utilizes previously created sequences for a series of training programmes in order to uncover optimizations as well as their deployment order. A clustering method selects optimizations during the first stage, and then a metaheuristic algorithm determines the order wherein the compiler would execute every optimization in the latter. The LLVM compilers as well as an I7 processor have been used to assess this strategy.

Supun et al. [17] developed HUMMINGBIRD, a unique prototype scoring technique that incorporates featurization operators with classic ML designs (e.g., decision trees) into a limited collection of tensor operational processes. This method decreases infrastructure overhead by using current investments in Neural Net compilers but also runtimes to produce efficient calculations for both CPU as well as hardware accelerators. The findings indicate that HUMMINGBIRD performs compatible.

Mircea et al. [18] introduced MLGO1, a methodology for comprehensively incorporating machine learning methods into an industrial compiler— LLVM. It's the first time ML has been fully integrated in a sophisticated compiler run in a real-world context. It's in the LLVM main repository. As contrasted to the state-of-the-art LLVM -Oz, we apply two alternative ML techniques to train the inlining-for-size method: Policy Gradient as well as Evolutionary Algorithms.

Aleksandar et al. [19] presented a revolutionary JIT compiler inlining approach that gradually investigates a program's call network and switches between inlining as well as optimizations. Three new heuristics have been developed to steer this inliner. Graal, a dynamic JIT compiler for the HotSpot JVM, was used to create this technique. Benchmarks such as Java DaCapo, Scalabench, and others were utilized to test the suggested algorithm.

Conventional systems to prediction model creation frequently employ a random selection search strategy, which can often lead to information redundancy. Moreover, the sample program gets exposed to a fixed number of repetitive measurements due to the influence of run-time disturbances. Unfortunately, if there are few sounds, the recurrent measurements will lead to a significant loss of iterative

compilation time overheads. Decreasing iterative compilation overheads and predicting an appropriate compiler optimization with less computational time and increased compiler performance was still challenging.

## III. PROPOSED COMPILER OPTIMIZATION PREDICTION MODEL

This proposed compiler optimization model comprises three working phases: model training, feature extraction [24, 25], as well as model exploitation (feature selection). First, the inputs were fed into the model training phase, which tries to match the right weights as well as bias to a learning algorithm [26, 27] in order to minimize a loss function throughout the validation range. The retrieved characteristics, such as static, dynamic, as well as improved entropy, were then transferred to the model exploitation phase [21], where the optimal features were chosen utilizing the improved chaos game optimization. These optimized features were given to Convolutional Neural Network for prediction of compiler optimization. The architecture of our improved compiler optimization prediction model is given in Fig. 1.

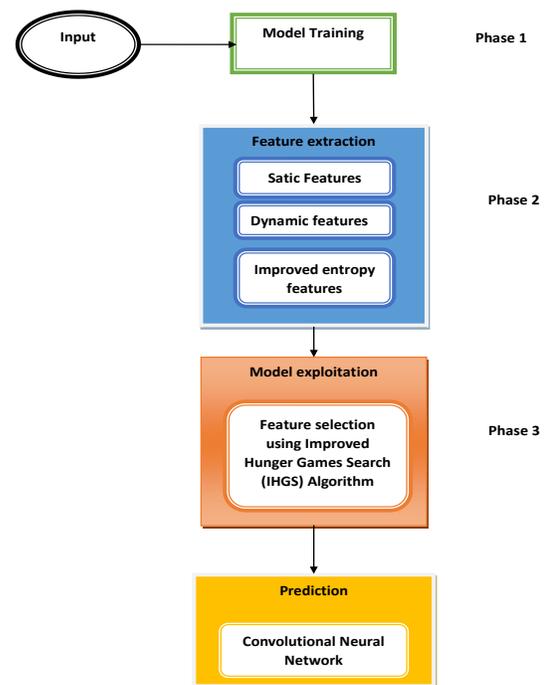


Fig. 1. Proposed improved compiler optimization prediction model architecture.

### A. Model Training Phase

The initialization and candidate sample set generation takes place in the model training phase. The initialization model will be built in the training set with some labeled samples. The initialization model will be used as the intermediate prediction model later. The candidate samples set include both the unlabeled samples in the training set and the labeled samples with the number of observations.

### B. Feature Extraction

Outputs from model training phase were given to the feature extraction phase to extract the static, dynamic and improved entropy features [26, 27].

1) *Static features*: The values of static features do not vary over time and are set for every sample. The lists of the static features extracted in this work were shown in Fig. 2.

2) *Dynamic features*: The values of dynamic features fluctuate over time and are not constant. Fig. 3 depicts the dynamic characteristics retrieved in this work.

3) *Improved entropy feature extraction*: The count of coefficients is generally so large that it is challenging to utilize them directly as features for categorization or prediction. As a result, several high-level features might emerge from these coefficients for improved prediction. Entropy seems to be a tool for measuring the uncertainty of data content in specific mechanisms, and it is frequently employed in signal analysis, pattern recognition, pattern matching, and other fields. . Some kinds of entropy include Shannon entropy (SE), log energy entropy (LEE), Renyi entropy (RE), as well as Tsallis entropy (TE). Renyi entropy is utilized to retrieve features from input data in this work. Entropy may be estimated via energy. Wavelet energy, described as Eq. (1), will be used to assess the data of the coefficient  $a$  of the  $b$ -th node at the  $c$ -th level.

$$E_{a,b,c} = \|d_{a,b,c}\|^2 \quad (1)$$

The total energy for the  $b$ -th node at the  $c$ -th level may then be determined utilizing Eq. (2)

$$E_{a,b} = \sum_{c=1}^M E_{a,b,c} \quad (2)$$

Where  $M$  indicates the number of node matching coefficients Eq. (3) may be used to compute the probability of the  $c$ -th coefficient at its associated node:

$$\rho_{a,b,c} = E_{a,b,c} / E_{a,b} \quad (3)$$

Where the sum of  $\rho_{a,b,c}$  equals 1.

Renyi Entropy of order  $q$  ( $q \geq 0$  and  $q \neq 1$ ) gets described as

$$RE(s) = \frac{1}{1-\beta} \log \left( \sum_{a=1}^M \rho_a^q \right) \quad (4)$$

The parameter of  $q$  in RE should be optimized to provide better results. In our work, the improved entropy features were extracted using the eq. (5)

$$RE(\rho) = \frac{1}{1-\beta} \log \left( \sum_{a=1}^M \rho_a^\beta \right)^{1/\beta} * \omega_a \quad (5)$$

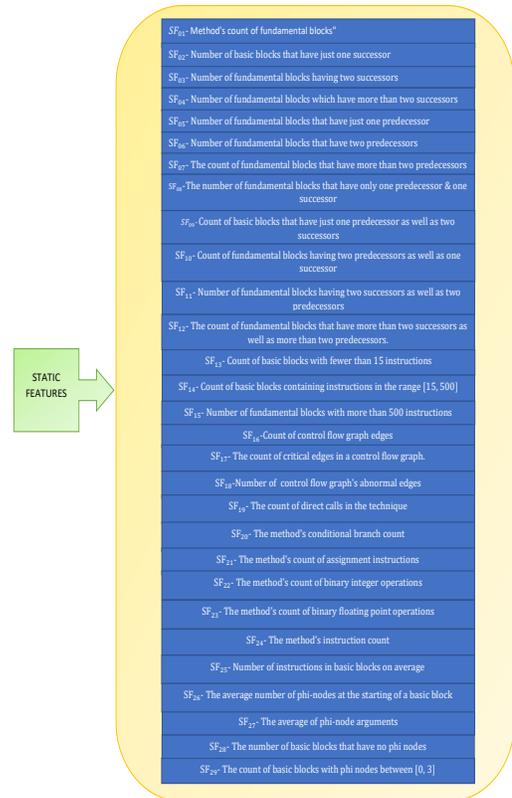


Fig. 2. Extracted static features.

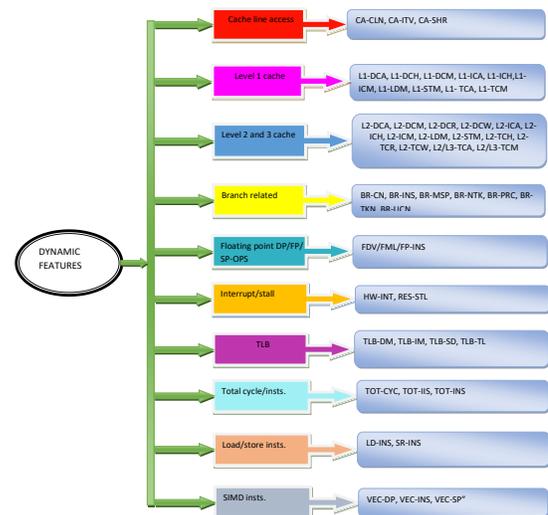


Fig. 3. Extracted dynamic features.

Here  $\omega_a$  denotes the weight, which is calculated by

$$\omega_a = \frac{\rho_a^\beta}{\sum_{b=1}^m \rho_b^\beta} \quad (6)$$

Following the calculation of the entropy of each terminal node, the entropies of all terminal nodes are concatenated to form a feature vector. These features were sent to the model exploitation phase for feature selection.

### C. Model Exploitation

The feature selection procedure is taking place during this model exploitation phase. When creating forecasting models, feature selection is the technique of minimizing the count of input variables. It is preferable to limit the count of input variables in order to reduce modeling computational costs and, in certain situations, increase model performance. For that reason we used an improved HGS Optimization.

1) *Improved Hunger Game Search (IHGS) Optimization:* The HGSO technique is influenced by normal animal behaviors including terror of being eaten by predators as well as hunger. The mathematical modelling of the HGSO strategy is explained in this portion of the publication. The modelling is based on social selection and hunger-driven behaviour.

This section quantitatively models the approaching behaviour of hunger. Eq. (7), which explains the foraging hunger as well as individual supportive communication activities, contains the game instructions. The contraction mode is imitated by the mathematical formula in Eq. (7).

$$\overline{Z}(t+1) = \begin{cases} \overline{Z}(t) \cdot (1 + \text{randn}(1)), & r1 < l \\ \overline{W}_1 \cdot \overline{Z}_d + \overline{R} \cdot \overline{W}_2 \cdot |\overline{Z}_d - \overline{Z}(t)|, & r1 > l, r2 > H \\ \overline{W}_1 \cdot \overline{Z}_d - \overline{R} \cdot \overline{W}_2 \cdot |\overline{Z}_d - \overline{Z}(t)|, & r1 > l, r2 < H \end{cases} \quad (7)$$

where  $\overline{Z}(t)$  signifies the location of all individuals,  $\overline{Z}_d$  indicates the position of the best individual,  $\overline{W}_1$  and  $\overline{W}_2$

seem to be hunger weights of hunger,  $\overline{R}$  is between [-a, a], r1 and r2 seem to be random numbers between [0, 1], randn(1) denotes a normal distributed random number, and t seems to be the count of current iterations. The parameter l represents the HGSO algorithm's control variable that governs the algorithm's sensitivity. H stands for variation control for all locations.

2) *Opposite behavior learning:* Amongst the most effective instructional procedures, opposition-based learning (OBL), has been extensively embraced as an excellent learning phase to improve the searching capabilities of algorithms. When assessing a solution Y to a given issue, a novel opportunity will be gained that brings the candidate solution closer to the optimal solution if the opposing solution of Y is estimated at the same time. The opposing number as well as opposite point notions were described as follows.

OBL is a learning approach that is centered on the inverse number  $Y^o$ . Y is described as a real number,  $Y \in [e, f]$ .  $Y^o$ 's opposite may be defined as (8), where e, f are the bounds.

$$Y^o = e + f - Y \quad (8)$$

When  $Y = (Y_1, Y_2, \dots, Y_D)$  seems to be a point in a D-dimensional space,  $Y_j, \dots, Y_D$ .  $e_j$  as well as  $f_j$  represent the current population's low and high borders, which vary with

each iteration. An opposing point in several dimensions has been described as

$$Y_j^o = e_j + f_j - Y_j, j = 1: D \quad (9)$$

In our work, to generate chaotic opposite solution we have used the following equation

$$Y_j = lbj + \text{rand} * (efj - lfj) \quad (10)$$

Here rand was generated using the sine map.

$$z_{u+1} = \frac{e}{4} \sin(\pi z_k) \quad e \in (0, 4) \quad (11)$$

Variation control for all positions H stated in eq. (12)

$$H = \text{sech}(|F(i) - BF|) \quad (12)$$

where F(i) represents the cost function value of every population,  $i = 1, 2, \dots, n$ , BF represents the best cost function value acquired during the latest incarnation, and Sech represents the hyperbolic function and thus is equal to  $(\text{sech}(x) = \frac{2}{e^x + e^{-x}})$ .

In our work, we used the reciprocal of the hyperbolic function Csch, which is expressed in eqn., (13)

$$H = \text{Csch}(|F(k) - BF|) \quad (13)$$

$$\text{Here Csch} = \frac{2e^x}{2e^x - 1}$$

Eqn., (14) gives the expression for  $\overline{R}$ .

$$\begin{aligned} \overline{R} &= 2 \times h \times r - h \\ h &= 2 \times \left(1 - \frac{t}{\text{Max}_{iter}}\right) \end{aligned} \quad (14)$$

Where  $\text{Max}_{iter}$  denotes the maximum number of iterations and rand denotes a random number between [0, 1]

3) *Hunger role:* This portion quantitatively models the hunger behavior of all individuals during the search. The formula for  $\overline{W}_1$  is given in Eq. (15).

$$\overline{W}_1(k) = \begin{cases} \text{hungry}(k) \cdot \frac{N}{SHungry} \times ra4, & ra3 < l \\ 1 & ra3 > l \end{cases} \quad (15)$$

The expression for  $\overline{W}_2$  is presented in Eq. (16).

$$\overline{W}_2(k) = (1 - \exp(-|hungry(k) - SHungry|)) \times ra5 \times 2 \quad (16)$$

where N represents population size, hungry means population starvation, SHungry represents the total of population starvation, i.e., sum(hungry), as well as ra3, ra4, and ra5 signify random values between [0, 1]. Each population's starving is quantitatively represented in Eq. (17).

$$\text{hungry}(k) = \begin{cases} 0 & \text{AllFitness}(k) == BF \\ \text{hungry}(k) + H_{new}, & \text{AllFitness}(k) = BF \end{cases} \quad (17)$$

where  $AllFitness(k)$  would be the present iteration's cost function value for every population. Depending on the real starving, a new starvation  $H_{new}$  is added. The equation represents the formula for  $H_{new}$ .

$$H_{new} = \begin{cases} LH \times (1 + ra) & TH < LH \\ TH, & TH \geq LH \end{cases} \quad (18)$$

$$TH = \frac{F(k) - BF}{WF - BF} \times ra \times 2 \times (UB - LB) \quad (19)$$

where  $H_{new}$  has been constrained to a lower bound LH, ra would be a random number among [0, 1], WF as well as BF seem to be the worst best fitness acquired during the latest iteration, respectively, F (k) has become the fitness of every population, ra would be a random number between [0, 1], and LB as well as UB are indeed the lower and upper boundaries of the dimensions, respectively. The selected features were sent to the CNN for compiler optimization prediction.

#### D. Prediction using CNN

Convolutional networks were deep training strategies that extract information from input pictures by convolving them with filters or kernels. Convolution of a GCG picture with a  $f_s C f_s$  filter learns the same characteristic on the whole picture. After each action, the window moves, and the feature maps learn the characteristics. The feature maps record the image's local receptive area and operate with mutual weights as well as biases. Equation (20) depicts the output matrix size without padding, whereas Equation (21) depicts the convolution procedure. Padding has been utilized to keep the size of the given picture constant. The output picture size is the same as the input image size in a 'SAME' padding, and there is no padding in a "VALID" padding. Equation depicts the output matrix size with padding (22).

$$GCG * f_s C f_s = G - F + 1 \quad (20)$$

$$o = \sigma \left( m + \sum_{v=0}^2 \sum_{q=0}^2 w_{v,q} \chi_{\alpha+v,m+q} \right) \quad (21)$$

$$GCG * f_s * f_s = (G + 2P - f_s) / (S + 1) \quad (22)$$

Here, O is the output, P is the padding, S is the stride, m is the bias,  $\sigma$  is the sigmoidal activation function, w is a 3x3 weight matrix of shared weights and  $\chi_{p1,p2}$  is the input activation at position p1,p2. The output O provides the prediction results.

## IV. RESULTS AND DISCUSSION

### A. Simulation Setup

The unique methodology for compiler optimization utilizing IHGS was implemented in Python. The standard performance evaluation group created the SPEC CPU2006 training set to evaluate general-purpose CPU performance [20].

The input scale of the SPEC2006 benchmark may be split into test, train, as well as reference scales; we utilize the reference scale to test." In this case, analysis was performed for multiple measures such as accuracy [22,23] and error metrics such as MSE, MSLE, and so on. In addition, IGHS outperformed the HGS, PRO, CMBO, ARCHOA, DO, as well as GOA models.

### B. Performance Analysis

The research on diverse metrics including accuracy, sensitivity, specificity as well as precision was detailed here. Here, the analysis was done for LPs (Learning Percentages) of 60, 70, 80 and 90 over HGS, PRO, CMBO, ARCHOA, DO, GOA models which is shown in Fig. 4. For 60 LP, CMBO and HGS achieve the accuracy rate of 0.69 and 0.76 whereas our proposed IHGS model achieves the accuracy rate of 0.84. At 80 and 90 LPs our proposed IHGS achieves the accuracy rate of 0.9 and 0.94 which is higher than other models. When our proposed IHGS achieves the precision value of 0.9, ARCHOA, CMBO models achieves only 0.8 and 0.81 for 60 LP which proves the superiority of proposed IHGS model. For 80 and 90 LPs, PRO model attain the sensitivity and specificity values of 0.83, 0.85 and 0.83, 0.85 while our proposed IHGS method achieves the values of 0.85, 0.93 and 0.89, 0.94 which proves that our proposed IHGS method achieves high performance for the compiler optimization identification than other conventional models.

The most often employed KPIs to estimate forecast accuracy were MAPE, MAE, RMSE(MSE), as well as MSLE which were analyzed for the models such as HGS, PRO, CMBO, ARCHOA, DO and GOA for 60, 70, 80 and 90 LPs which is compared with our proposed IHGS model that is shown in Fig. 5. "MAE is indeed a metric of error between matched observations reflecting the same phenomena in statistics." The MAE should be less to increase forecast accuracy. Our proposed IHGS method obtain the MAE value of 0.48, 0.45, 0.42 and 0.4 for 60, 70, 80 and 90 LPs which is lower than other conventional methods. MSLE may be regarded of as a measurement of the ratio between true as well as forecasted values. When HGS method achieves the high MAPE values of 2.3, 1.5, 0.7 and 2.0 for 60, 70, 80,90 LPs, our proposed IHGS method obtain the values of 0.5, 0.4, 0.3 and 0.2. Unlike MAE, RMSE doesn't really handle every error in the same way. It prioritises the most critical errors. That implies that a single large mistake might result in a very bad RMSE. Our proposed IHGS approach yields MSE values of 0.49, 0.46, 0.43, as well as 0.42 for all LPs, which is lower than other standard approaches. In statistics, the MAPE, also referred as the MAPD, is specified as "a metric of prognosis accuracy of a forecasting technique". "The MSE or MSD of an estimator in statistics estimates the average of the squares of the errors, or the average squared difference between the predicted as well as real values." For optimized prediction, the MSE and MAPE must be lower. When the CMBO approach produces MSLE values of 0.27, 0.22, 0.23, 0.26, our proposed IHGS method achieves lower values of 0.23, 0.22, 0.21, 0.20, demonstrating that our proposed IHGS method can outperform other standard compiler optimization forecasting models.

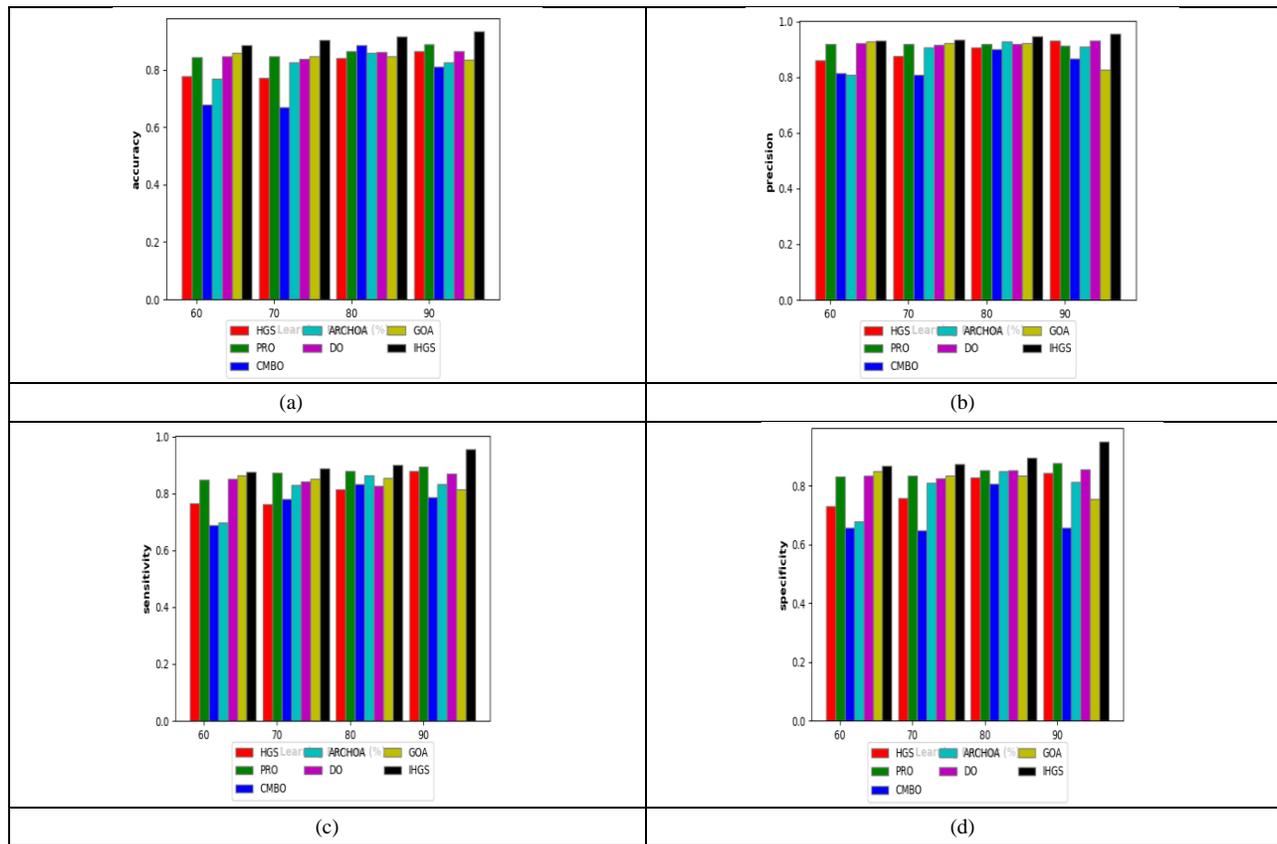


Fig. 4. Comparison of performance matrices such as (a)Accuracy, (b)Precision, (c) Sensitivity, (d) Specificity.

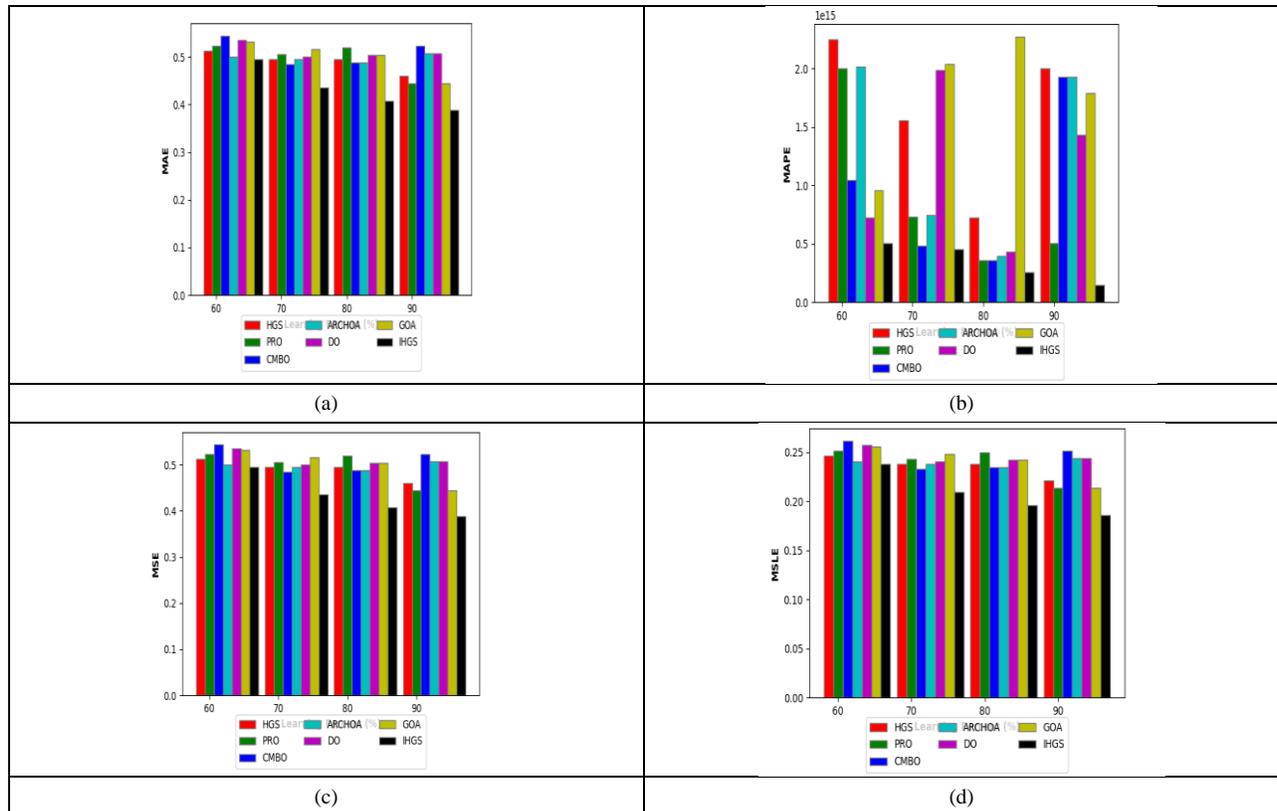


Fig. 5. Comparison of performance such as (a) MAE, (b) MAPE, (c) MSE, (d) MSLE.

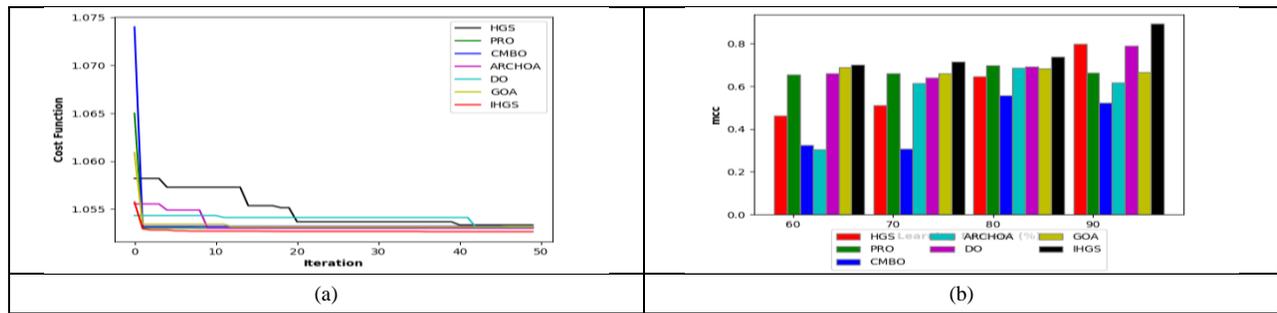


Fig. 6. Comparison of MSC and Cost function values for different LPs.

The cost function for 0-50 iterations was evaluated in this work, to assess the performance of our proposed IHGS model which is shown in Fig. 6(a). When cost function of CMBO is 1.074, our proposed method obtains the value of 1.052 for iteration 0 whereas 1.058 and 1.062 for GOA method which states that IHGS method has the lowest cost function for all five iterations. A more trustworthy statistical rate known as the Matthews correlation coefficient (MCC) was evaluated, which yields a high score only if the prediction performed well in all the confusion matrix classes which are given in Fig. 6(b). MCC values of all the LPs were 0.67, 0.68, 0.7 and 0.9 for our proposed IHGS method which proves our prediction was performed well with good results.

The F-measure is derived as the harmonic mean of accuracy as well as recall, with equal weighting for each. It

enables a system to be assessed utilizing a single score that accounts for both accuracy and recall that is useful for reporting system performance as well as comparing models. With F1 measure, fnr, fpr, as well as npv values were also estimated and compared with conventional models which is shown in Fig. 7. In comparison to the CMBO as well as ARCHOA approaches, our proposed IHGS method achieves f1 measure values of 0.9, 0.92, 0.93, and 0.95 for all LPs. The IHGS approach produces npv value of 0.92 for 90 LP, whereas the CMBO and ARCHOA methods yield relatively low values such as 0.6 and 0.68. Our proposed IHGS technique achieves fpr values of 0.13, 0.12, 0.11, and 0.05, and assessed fnr values of 0.13, 0.12, 0.11, and 0.04, which are lower than other traditional methods, demonstrating that IHGS method achieves superior performance than other methods.

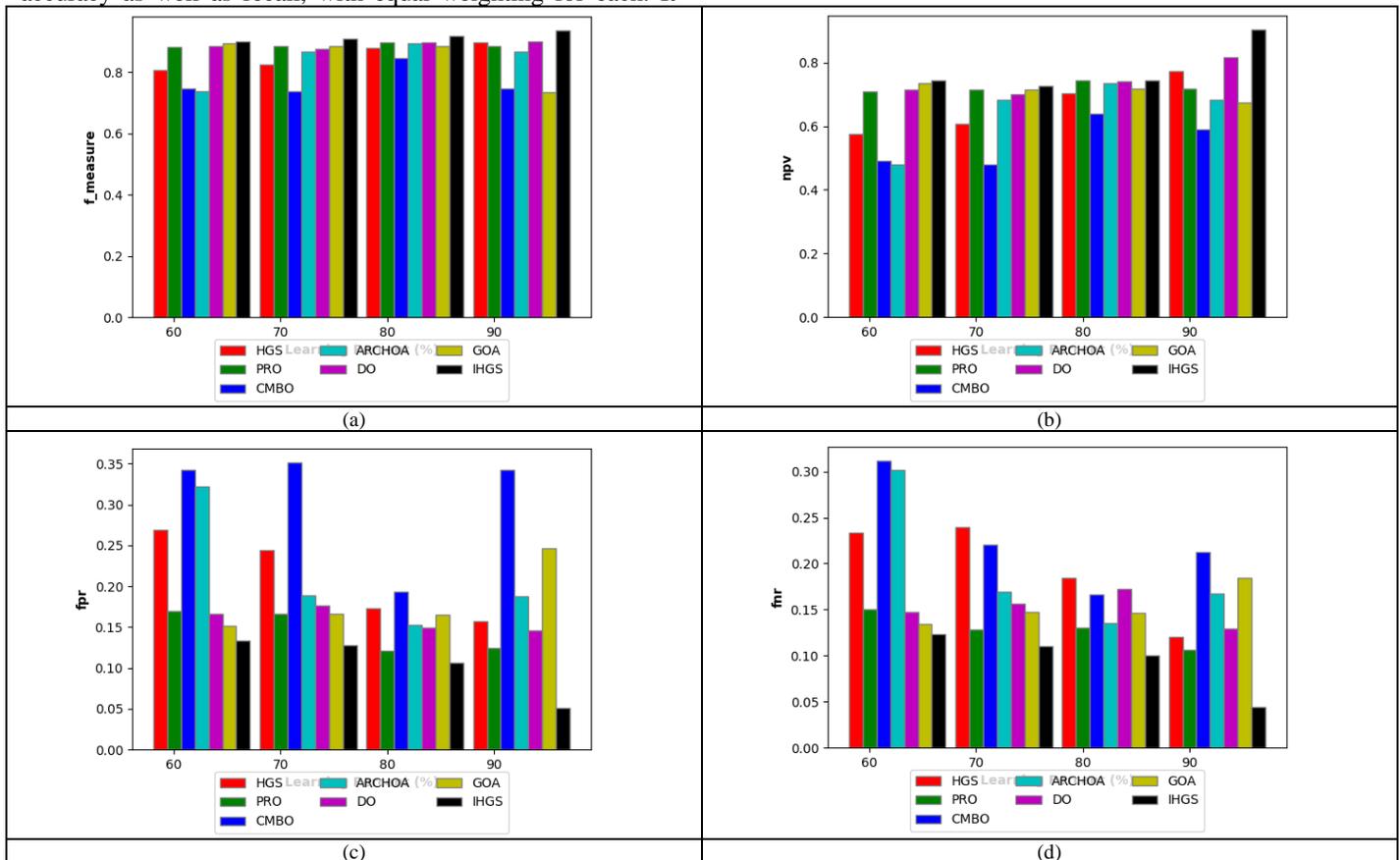


Fig. 7. Comparison of performance matrices (a) F1 measure, (b) fnr, (c) fpr, (d) npv values.

The computation time for each method which is compared with our proposed IGHS method is shown in Table I which shows that IGHS method have low computational time of 55.64. Accuracy and error matrices also compared with without optimization, to evaluate the performance of our proposed IGHS method which is given in Tables II and III. Without optimization our proposed method achieves only 88% accuracy whereas, with optimization it achieves 95% accuracy.

TABLE I. COMPUTATIONAL TIME COMPARISON FOR CONVENTIONAL AND PROPOSED APPROACHES

Methods	Computation time
HGS	76.2114
PRO	200.538
CMBO	537.269
ARCHOA	139.731
DO	95.0285
GOA	111.956
IHGS	55.6467

TABLE II. PERFORMANCE MATRICES OF THE PROPOSED IGHS METHOD (ACCURACY MATRICES) WITH AND WITHOUT OPTIMIZATION

Accuracy matrices	Proposed with Optimization	Proposed without Optimization
sensitivity	0.955614	0.882424
specificity	0.949348	0.765633
accuracy	0.933705	0.843407
precision	0.955031	0.882424
F -measure	0.936337	0.882424
mcc	0.892841	0.648057
npv	0.903579	0.765633
fpr	0.050652	0.234367
fnr	0.044386	0.117576

TABLE III. PERFORMANCE MATRICES OF THE PROPOSED IGHS METHOD (ERROR MATRICES) WITH AND WITHOUT OPTIMIZATION

Error matrices	Proposed with Optimization	Proposed without Optimization
MSE	0.388571	0.515957
MAE	0.388571	0.515957
MSLE	0.185908	0.247893
MAPE	1.43E+14	2.32E+15

Table IV Shows the RMSE values obtained for distinct datasets and statistical tests such as Wilcoxon and chi-square were conducted for conventional and proposed methods and the p, statistic values were tabulated in Tables V and VI which proves the effectiveness of our proposed compiler optimization prediction approach.

TABLE IV. RMSE FOR EACH BENCHMARK IN THE DATASET

Bench mark	RMSE
400.perlbenc	0.701
401.bzip2	0.707107
403.gcc	0.701
429.mcf	6.61E-01
445.gobmk	0.809156
456.hmmer	0.75
458.sjeng	0.75
462.libquantum	0.707107
464.h264ref	0.661438
471.omnetpp	0.696107
473.astar	0.612372
483.xalancbmk	0.644378

TABLE V. COMPARISON OF WILCOXAN TEST RESULTS FOR PROPOSED AND CONVENTIONAL METHODS

Methods	P value	Statistic
HGS	1.36E-06	253
PRO	6.48E-18	2701
CMBO	2.54E-29	7475
ARCHOA	6.97E-29	7.63E+03
DO	3.55E-12	1128
GOA	2.54E-29	7475
IHGS	2.54E-29	7875

TABLE VI. CHI-SQUARE TEST RESULTS FOR PROPOSED AND TRADITIONAL TECHNIQUES

Methods	P value	Statistic
HGS	5.83E-03	45
PRO	8.83E-05	59
CMBO	5.83E-03	45
ARCHOA	5.83E-03	4.50E+01
DO	4.37E-06	68
GOA	5.83E-03	45
IHGS	1.54E-06	71

## V. CONCLUSION

Selecting the optimal, or even a good, combination of optimizations for an unpredictable programmed on an arbitrary design is a task so tough that traditional manual analysis approaches are impractical. For that reason, a novel optimization prediction model with improved optimization was developed in this work, which has three working phases including model training, feature selection as well as feature exploitation phase. First, the inputs are being sent to the model

training phase that aims to link the appropriate weights as well as bias to a learning algorithm in order to minimize a loss function throughout the validation range. The retrieved characteristics, including static, dynamic, and enhanced entropy, were then transferred to the model exploitation phase, where the best features was determined using the improved chaos game optimization. These improved characteristics were fed into a Convolutional Neural Network to predict the appropriate compiler optimization.

## REFERENCES

- [1] Ashouri, A.H., Killian, W., Cavazos, J., Palermo, G. and Silvano, C., 2018. A survey on compiler autotuning using machine learning. *ACM Computing Surveys (CSUR)*, 51(5), pp.1-42.
- [2] Georgiou, K., Chamski, Z., Amaya Garcia, A., May, D. and Eder, K., 2022. Lost in translation: Exposing hidden compiler optimization opportunities. *The Computer Journal*, 65(3), pp.718-735.
- [3] Wang, Z. and O'Boyle, M., 2018. Machine learning in compiler optimization. *Proceedings of the IEEE*, 106(11), pp.1879-1901.
- [4] Tağtekin, B., Höke, B., Sezer, M.K. and Öztürk, M.U., 2021, August. FOGA: Flag Optimization with Genetic Algorithm. In *2021 International Conference on INnovations in Intelligent SysTems and Applications (INISTA)* (pp. 1-6). IEEE.
- [5] Gong, Z., Chen, Z., Szaday, J., Wong, D., Sura, Z., Watkinson, N., Maleki, S., Padua, D., Veidenbaum, A., Nicolau, A. and Torrellas, J., 2018. An empirical study of the effect of source-level loop transformations on compiler stability. *Proceedings of the ACM on Programming Languages*, 2(OOPSLA), pp.1-29.
- [6] Chen, J., Xu, N., Chen, P. and Zhang, H., 2021, May. Efficient compiler autotuning via bayesian optimization. In *2021 IEEE/ACM 43rd International Conference on Software Engineering (ICSE)* (pp. 1198-1209). IEEE.
- [7] L. P. Cáceres, F. Pagnozzi, A. Franzin, and T. Stütze, "Automatic configuration of gcc using irace," in *International Conference on Artificial Evolution (Evolution Artificielle)*. Springer, 2017, pp. 202–216.
- [8] Phothilimthana, P.M., Sabne, A., Sarda, N., Murthy, K.S., Zhou, Y., Angermueller, C., Burrows, M., Roy, S., Mandke, K., Farahani, R. and Wang, Y.E., 2021, September. A Flexible Approach to Autotuning Multi-Pass Machine Learning Compilers. In *2021 30th International Conference on Parallel Architectures and Compilation Techniques (PACT)* (pp. 1-16). IEEE.
- [9] Chen, T., Moreau, T., Jiang, Z., Zheng, L., Yan, E., Shen, H., Cowan, M., Wang, L., Hu, Y., Ceze, L. and Guestrin, C., 2018. {TVM}: An Automated {End-to-End} Optimizing Compiler for Deep Learning. In *13th USENIX Symposium on Operating Systems Design and Implementation (OSDI 18)* (pp. 578-594).
- [10] Steiner, B., Cummins, C., He, H. and Leather, H., 2021. Value learning for throughput optimization of deep learning workloads. *Proceedings of Machine Learning and Systems*, 3, pp.323-334.
- [11] Zheng, L., Jia, C., Sun, M., Wu, Z., Yu, C.H., Haj-Ali, A., Wang, Y., Yang, J., Zhuo, D., Sen, K. and Gonzalez, J.E., 2020. Anso: Generating {High-Performance} Tensor Programs for Deep Learning. In *14th USENIX Symposium on Operating Systems Design and Implementation (OSDI 20)* (pp. 863-879).
- [12] Jia, Z., Padon, O., Thomas, J., Warszawski, T., Zaharia, M. and Aiken, A., 2019, October. TASO: optimizing deep learning computation with automatic generation of graph substitutions. In *Proceedings of the 27th ACM Symposium on Operating Systems Principles* (pp. 47-62).
- [13] Zheng, S., Liang, Y., Wang, S., Chen, R. and Sheng, K., 2020, March. Flextensor: An automatic schedule exploration and optimization framework for tensor computation on heterogeneous system. In *Proceedings of the Twenty-Fifth International Conference on Li, M., Zhang, M., Wang, C. and Li, M., 2020. Adatune: Adaptive tensor program compilation made efficient. Advances in Neural Information Processing Systems*, 33, pp.14807-14819.
- [14] Li, M., Zhang, M., Wang, C. and Li, M., 2020. Adatune: Adaptive tensor program compilation made efficient. *Advances in Neural Information Processing Systems*, 33, pp.14807-14819.
- [15] Liu, H., Zhao, R., Wang, Q. and Li, Y., 2018. ALIC: A low overhead compiler optimization prediction model. *Wireless Personal Communications*, 103(1), pp.809-829.
- [16] de Souza Xavier, T.C. and da Silva, A.F., 2018. Exploration of compiler optimization sequences using a hybrid approach. *Computing and Informatics*, 37(1), pp.165-185.
- [17] Nakandala, S., Saur, K., Yu, G.I., Karanasos, K., Curino, C., Weimer, M. and Interlandi, M., 2020. A tensor compiler for unified machine learning prediction serving. In *14th USENIX Symposium on Operating Systems Design and Implementation (OSDI 20)* (pp. 899-917).
- [18] Trofin, M., Qian, Y., Brevdo, E., Lin, Z., Choromanski, K. and Li, D., 2021. Mlgo: a machine learning guided compiler optimizations framework. *arXiv preprint arXiv:2101.04808*.
- [19] Prokopec, A., Duboscq, G., Leopoldseder, D. and Wirthinger, T., 2019, February. An optimization-driven incremental inline substitution algorithm for just-in-time compilers. In *2019 IEEE/ACM International Symposium on Code Generation and Optimization (CGO)* (pp. 164-179). IEEE Algorithm for just in time compilers. In *2019, IEEE/ACM International Symposium on Code Generation and Optimization(CGO)*(pp.164-179).
- [20] SPEC CPU2006: SPEC CPU2006 benchmark suite. <http://www.spec.org/cpu/>.
- [21] Sandeep U. Kadam, Sagar B. Shinde, Yogesh B. Gurav, Sunil B Dambhare and Chaitali R Shewale, "A Novel Prediction Model for Compiler Optimization with Hybrid Meta-Heuristic Optimization Algorithm" *International Journal of Advanced Computer Science and Applications(IJACSA)*, 13(10), 2022. <http://dx.doi.org/10.14569/IJACSA.2022.0131068>.
- [22] A. D. Sutar and S. B. Shinde, "ECU diagnostics validator using CANUSB," 2017 International Conference on Inventive Computing and Informatics (ICICI), Coimbatore, India, 2017, pp. 856-860, doi: 10.1109/ICICI.2017.8365257.
- [23] A. D. Sutar and S. B. Shinde, "ECU Health Monitor Using CANUSB," 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT), Coimbatore, India, 2018, pp. 415-419, doi: 10.1109/ICICCT.2018.8473000.
- [24] S. Shinde and R. B. Waghulade, "An improved algorithm for recognizing mathematical equations by using machine learning approach and hybrid feature extraction technique," 2017 IEEE International Conference on Electrical, Instrumentation and Communication Engineering (ICEICE), Karur, India, 2017, pp. 1-7, doi: 10.1109/ICEICE.2017.8191926.
- [25] S. Shinde, R. B. Waghulade and D. S. Bormane, "A new neural network based algorithm for identifying handwritten mathematical equations," 2017 International Conference on Trends in Electronics and Informatics (ICEI), Tirunelveli, India, 2017, pp. 204-209, doi: 10.1109/ICOEI.2017.8300916.
- [26] Kalyani Wagh, K. Vasanth, Sagar Shinde , "Emotion Recognition Based On Eeg Features With Various Brain Regions", *Indian Journal of Computer Science and Engineering (IJCSSE)*, Vol. 13 No. 1 Jan-Feb 2022, p-ISSN : 2231-3850, DOI : 10.21817/indjcsse/2022/v13i1/221301095.
- [27] Khoje, s., Shinde, S. Evaluation of Ripplet Transform as a Texture Characterization for Iris Recognition. *J. Inst. Eng. India Ser. B* (2023). <https://doi.org/10.1007/s40031-023-00863-6>.

# Design of an English Web-based Teaching Resource Sharing Platform based on Mobile Web Technology

Yan Zhang

Huanghe Jiaotong University, Basic Education Department, Jiaozuo, 454950, China

**Abstract**—Thanks to the booming technology of computers and multimedia, student-centered online teaching resource platforms have become an important way for students to learn. However, English teaching resource platforms at the present stage fail to effectively integrate the massive and scattered learning resources. Based on this, the study proposes an English online teaching resource sharing platform based on mobile Web technology, using the SOAP protocol to deploy heterogeneous data resources as Web services to achieve interchangeability between heterogeneous resources. In addition, to enhance the efficient use of learning resources by students, the study proposes a hybrid algorithm based on collaborative filtering algorithm and sequential pattern mining algorithm to achieve personalized sequential recommendation for students. The results show that the platform created by the study exhibits excellent performance in terms of resource transfer capability, achieves efficient teaching resource sharing in a short response time and also shows that the proposed recommendation algorithm is highly accurate.

**Keywords**—Web service; SOAP; English teaching resources; sharing platform; personalised recommendation

## I INTRODUCTION

With the continuous promotion of education reform, the value of online education has begun to emerge, which has a positive impact on the development of education in all disciplines [1]. English, as a basic language discipline, in the process of combining with online education, in addition to its own development towards diversification and multidisciplinary integration, it also acts as the language carrier of other disciplines [2]. Therefore, English teaching research is particularly important. At present, many colleges and universities are organizing English teachers and educational technicians to integrate teaching resources, develop online courses, provide students with a good internal environment, and realize the sharing of teaching resources [3]. However, the current multimedia resource sharing platform has caused the problem of excessive platform resources due to insufficient recommendation accuracy and low efficiency [4]. In this way, teachers need to carry out a lot of repeated work, and students and teachers' information acquisition ability is insufficient to improve teachers' teaching efficiency and students' learning autonomy [5]. In order to effectively integrate heterogeneous and dispersed English teaching resources and improve the quality of online English education, the research is based on mobile Web technology and resource recommendation optimization, and an online English teaching resource sharing platform is constructed.

The education platform is constructed by means of resource replacement, access protocol, user information encryption and algorithmic assistance. The MD5 algorithm is designed based on SOAP to ensure the user's information login security. In order to provide personalized learning services for students at different levels, a hybrid optimization algorithm recommendation model is proposed based on collaborative filtering algorithm and sequential pattern mining algorithm. The algorithm model can effectively improve the efficiency of recommendation, improve the accuracy of recommendation of English teaching resources, and better adapt to the current English education environment.

The research innovatively applies the combination of Web technology and English education resource recommendation method to the optimization of English education, providing a reference for the selection of the optimization direction of multi-class online education. In addition, the sequence mining method is incorporated into the collaborative recommendation algorithm, and the comprehensive efficiency of the recommendation algorithm is proposed, which provides a reference for the development of educational resource recommendation.

The research content mainly includes four parts. The second part is a summary of the research status of the construction of educational resources platform at home and abroad; The third part puts forward the construction of the English shared resources platform based on web services, Part A establishes the framework of the education resources platform, describes the construction of the MD5 encryption module, Part B constructs the online learning resources recommendation framework based on web logs, and specifically analyzes the optimization of the recommendation algorithm; The fourth part verifies the application effect of the English online teaching resource sharing platform. The results show that the English online teaching resource sharing platform based on mobile Web technology and resource recommendation optimization has good application effect.

## II RELATED WORKS

In a web environment, resource allocation is a very important issue [6]. Yuan et al. proposed a new neural network path sorting algorithm based on path sorting after improving and analyzing the traditional algorithm, and used a path sorting based on random walk patterns and a neural network path sorting algorithm to solve the link prediction problem in online learning [7] Gu designed a recursive algorithm-led online training model using a web programming language as an example, and by testing the web programming language, it can

be seen that this training method has greatly helped students' learning effect [8]. To further improve the development of online courses, Zhang scholars developed a Mocoas-based mobile learning platform, and the experimental results showed that the platform can well meet the multiple teaching and learning needs of learners [9]. Fang scholars developed an online teaching assessment system based on machine learning, which can effectively assess the quality of teaching and learning in schools, which has good application prospects [10]. To address the problem of schools only being able to teach remotely during the epidemic, Christianson designed a remote web poll to improve students' immediate engagement in a virtual environment, which experimentally proved to be positively experienced by students [11]. Scholars such as Zhou used a WEB web-based teaching platform, using principal meta-analysis and clustering algorithm to classify students' performance and evaluate their learning effectiveness, and the experiment proved that this function can effectively help teachers improve their teaching [12]. Zhang's team attempted to use blockchain technology in the sharing of English teaching resources, and through the analysis of its algorithm, designed a representation layer, business layer and data layer as the core English teaching resources sharing platform, but the process of building the platform was tedious and consumed a lot of human and material resources [13]. Scholars such as Park S E proposed a resource sharing platform based on collaborative recommendation algorithm and introduced it to a hybrid recommendation-based system in recommending learning resources, but as the security of the platform was not taken into account, making the platform vulnerable to malicious attacks [14].

With the advent of the "Internet+" era, education informatics is placing increasing emphasis on the use of artificial intelligence technology to improve teaching effectiveness and quality. In intelligent teaching, intelligent recommendation is the key to realizing knowledge resources. After analyzing and comparing existing personalized recommendation technologies, Xu proposed a personalized recommendation algorithm based on content recommendation and collaborative screening, and the experimental results showed that the personalized recommendation algorithm has high correctness and effectiveness [15]. Based on the user interest model, Li et al. generated a user profile file by classifying the user's queries and generating a snapshot of the user's personal information through point-and-click, and the experimental results showed that the user profile was highly similar to the user's interests [16]. Chaabi's research group proposed a user interest model from a generalized to a specific hierarchy, which can effectively differentiate the interest characteristics of different classes [17]. Ohtomo et al. argue that users' interests when reading news can be divided into two categories, short term interests and long-term interests, with short term interests tending to be related to the timeliness of popular information and changing rapidly, while longer term interests tend to reflect users' real interests [18]. In contrast, Chen's team used a multi-strategy machine to build long-term and short-term interest patterns, and through an in-depth analysis of this system, a new content-based fusion algorithm was proposed and designed [19]. To improve the efficiency of the recommendation model, Wang combined the recommended

content with a collaborative filtering algorithm to provide a specific model for the actual and potential hybrid algorithm, and experimental results showed that the knowledge resource-based recommendation system was effective in improving the usability of the collaborative filtering part [20].

From the research of many scholars, it can be seen that the current research of online English education platform is still in the stage of rapid development, and the development of educational resources and the efficiency of the promotion of educational recommendation ability still do not match. In order to further improve the efficiency of recommendation and the utilization of resources, a resource sharing platform based on Web technology is designed. The collaborative filtering algorithm and sequential pattern mining algorithm are organically combined to form a new hybrid recommendation model, which aims to contribute to the research of English education resource recommendation.

### III WEB TECHNOLOGY PRACTICE FOR ENGLISH ONLINE TEACHING RESOURCE SHARING PLATFORM

#### A. Design of a Web Service-based English Shared Resource Platform

The intelligent network mobile teaching platform can provide learners with rich teaching resources and a student-oriented teaching environment. However, due to the wide distribution of network teaching resources, the diverse and complex structure of network resources between different campuses and the inability of data resources to cross the restrictions of platforms and firewalls, the resources between campuses cannot be shared effectively, resulting in the waste of campus teaching resources [21]. In order to solve the above problems, the study combines both sides of the teaching body with the trial module and builds an English resource sharing platform, using resource replacement - access protocols - user information encryption and algorithmic assistance to optimize the design of the platform system.

Service Oriented Architecture (SOA) is a component model that connects different functional units of services in an application by defining clear interfaces and contracts. web services are an effective way to implement SOA by abstracting applications and resources in a unified form through a standardized approach to service usage and by supporting It supports the sharing and collaboration of teaching and learning resources in a distributed environment. Among the technologies related to web services, web services are often used to solve integration problems across web applications due to their loosely coupled nature. Compared to traditional technologies, it not only solves the problem of incompatibility between heterogeneous systems, but also allows any user to make changes to the mechanism without affecting the normal operation of the platform. SOAP uses hypertext transfer protocol as the carrier form to achieve the construction of SOAP message structure model under the construction of each SOAP sub-element to ensure the integrity of SOAP messages, and the XML model as the encoding method for the exchange of resources, and its specific exchange process is shown in Fig. 1. The specific exchange process is illustrated in Fig. 1.

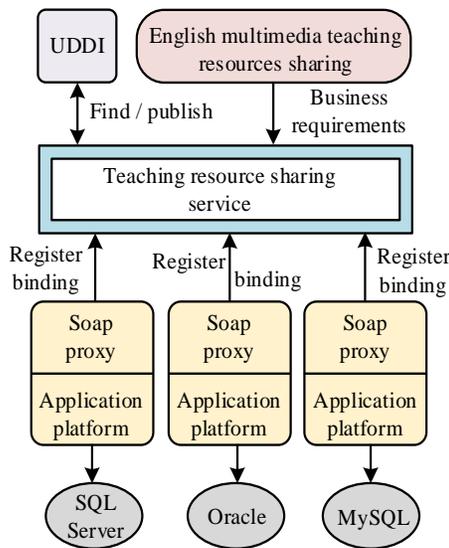


Fig.1. SOAP resource exchange process.

SOAP can effectively ensure that network service providers can realize the query of resource data, information update and sharing of different data types of each database. In order to further ensure the normal operation of the English multimedia resource platform and the confidentiality and security of user information data, the study is based on the MD5 information abstraction algorithm to accelerate the security of the platform information while ensuring high-quality services for users and realizing the optimal design of the platform security. The MD5 algorithm can effectively extend the length of the data and achieve a continuous distribution of bits at intervals of 512 times, thus forming a combination with regularity. Under the assumptions of the four MD5  $F(a,b,c)$ ,  $G(a,b,c)$ ,  $H(a,b,c)$ ,  $I(a,b,c)$  basis by bit can be expressed as eq. (1).

$$\begin{cases}
 F(a,b,c) = (a \& b) \mid ((\neg a) \& c) \\
 G(a,b,c) = (a \& c) \mid (b \& (\neg c)) \\
 H(a,b,c) = a \wedge b \wedge c \\
 I(a,b,c) = b \wedge (a \mid (\neg a))
 \end{cases}
 \tag{1}$$

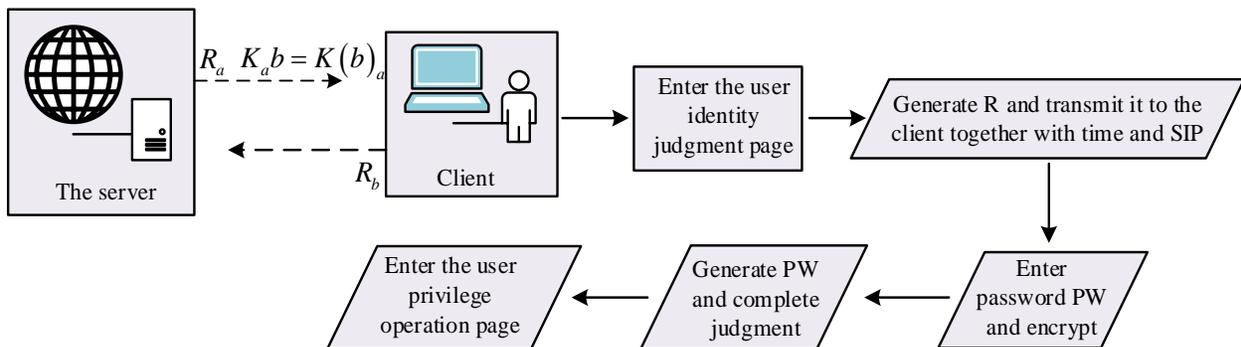


Fig.2. User identity verification process.

$Y_j$  The conversion of the grouping of  $(a,b,c)$  can be achieved in the form shown in eq. (2).

$$a_j = b + ((a + X(b,c,d) + Y_j + t_i) \lll s)
 \tag{2}$$

In the study the buffer corresponding to the MD5 algorithm is used as the link variable and the derivation of the message digest is implemented with the help of the buffer, which is represented as shown in equation (3). And after defining the four link variables, the weights corresponding to the variables are transformed with the help of equation (1) to realize the message data in order to achieve the confidentiality of the user information.

$$\begin{cases}
 A = 0X01234567 \\
 B = 0X89abcdef \\
 C = 0Xfecba98 \\
 D = 0X76543210
 \end{cases}
 \tag{3}$$

When the MD5 algorithm is used to verify user information on the resource sharing platform, it uses Ra random values to achieve circular encryption processing of the scheme, and increases the types and combinations of verification information data to reduce malicious attacks by illegal users and avoid leakage of user data, thus strengthening the overlay encryption of data, and its verification process is shown in Fig. 2.

When the information scheme is encrypted, in order to avoid the loss of resources under the platform due to the increased number of calibration repetitions, the server adds random values with immediate time characteristics to the string generation process by forming a completely new string from the random value R, the user password PW, the current time TIME and the address of the SIP connection protocol, as shown in eq. (4).

$$RA = R \mid TIME \mid SIP
 \tag{4}$$

The client also encrypts the plain text of the received password, as shown in eq. (5).

$$PW' = K(PW)
 \tag{5}$$

The server processes the generated sequence in depth using the MD5 algorithm to obtain the end-user's credentials and sends the MD data to the server, a process defined by eq. (6).

$$MD = MD5(R_a | PW') \quad (6)$$

The server application uses the MD5 algorithm to verify the encrypted text of the user's password and sends a confirmation message; the encrypted text is sent to the client and the passage of the user's credential information is used as a signal to achieve login. The research proposes a teaching resource platform that can effectively introduce SOAP into the construction of the platform to meet the demand characteristics of different users, on the basis of including the teaching sides module and the auditor module, and build a resource platform based on the cloud computing environment, whose overall structure as shown in Fig. 3.

As can be seen in Fig. 3, the software structure of the cloud computing environment as a platform for carrying content is divided into three layers: a data access layer, a business logic layer and a user performance layer. The data access layer enables the upload, audit and publication of learning materials and the download of matching resources with a list of completed learning materials, i.e. the resource download interface allows the setting of resource download links and binding them to the key values of the corresponding source files, providing users with an efficient resource download function. Finally, the audit resource fields and page call methods are retrieved by the teaching audit module to ensure that the resource platform can achieve good audit and coverage

of all learning resources and thus continuously enrich the types of teaching resources to meet the learning needs of students.

### B. Web Log-Based Online Learning Resource Recommendation Framework Design

As the amount of information carried on the Internet continues to grow, information referral methods have also flourished. In the vast amount of learning resources, it is difficult for learners to quickly find the information that meets their needs [22]. Therefore, the study proposes a hybrid recommendation algorithm based on the web-based learning platform logs by analyzing the logs of the above-mentioned web-based shared English learning platforms to uncover the characteristics of the resources and the preferences of users in the web environment. The hybrid recommendation algorithm combines a collaborative filtering algorithm and a sequential pattern mining algorithm in order to achieve a personalized recommendation model with real-time adaptability to the learning platform. The goal of collaborative filtering is to filter out data with similar characteristics from the database to form a single dataset, making the data as differentiated as possible between datasets based on different points of interest, while data belonging to the same dataset are as similar as possible. Assuming that each user in the learning platform for online educational resources  $X$  corresponds to a set of learning resources  $I_x$ , where the resource  $i \in I_x$ , which represents that the user has studied this resource  $i$ , has also been rated by the user, the user's scoring matrix for the resource is shown in Fig. 4(a).

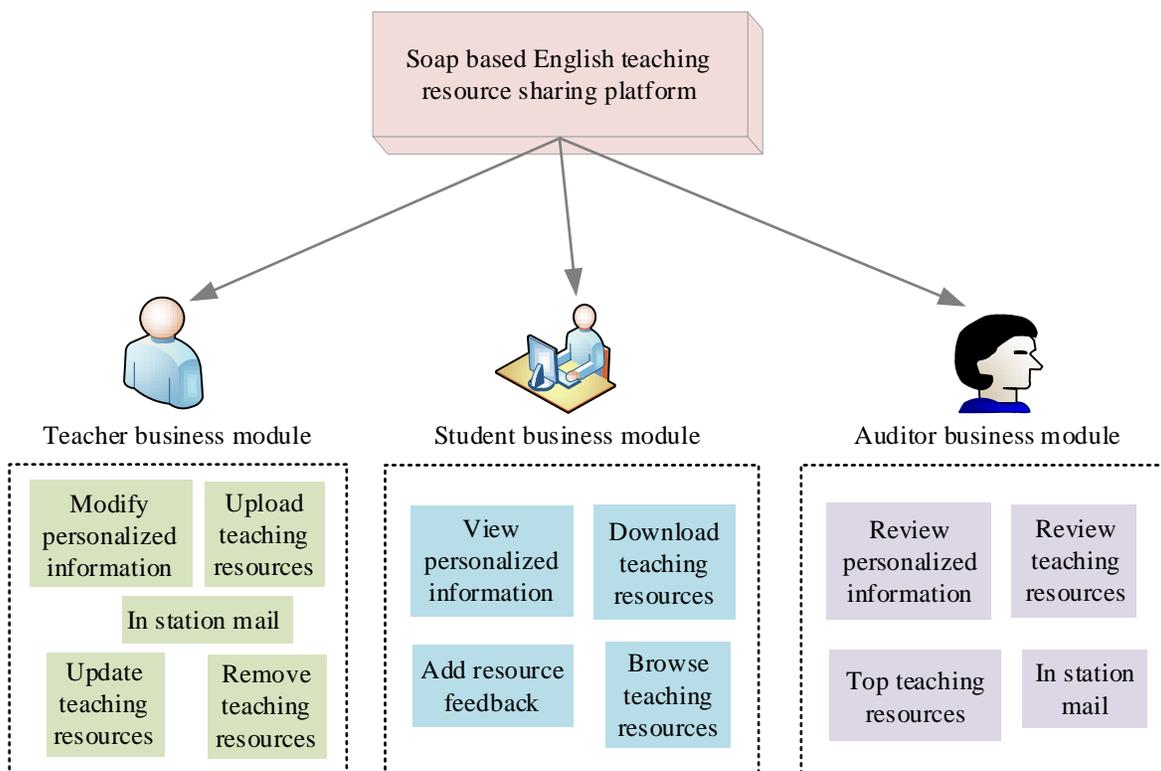


Fig.3. Schematic diagram of English multimedia teaching resource sharing platform based on SOAP.

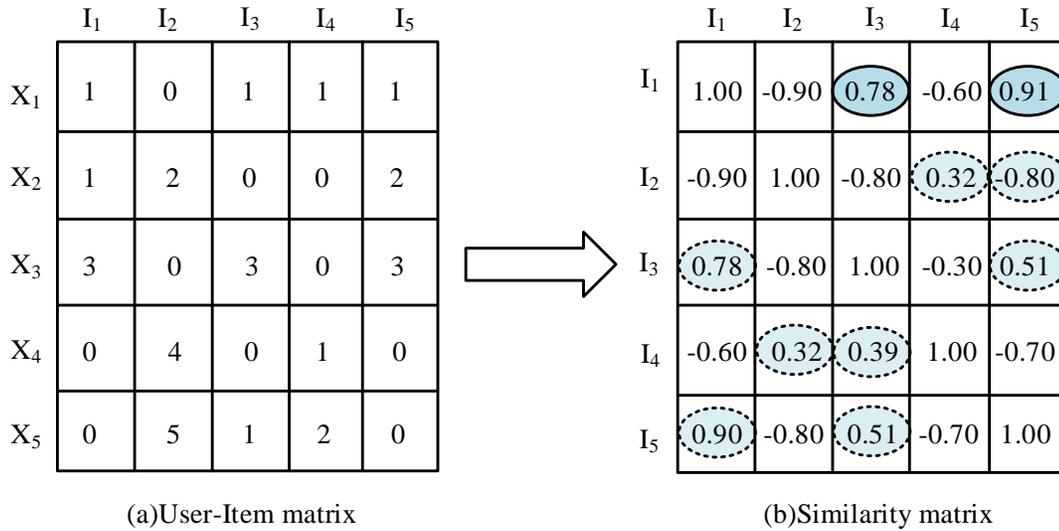


Fig.4. User resource rating and similarity matrix between resources.

Adjusted Cosine Similarity (ACS) is used to measure the similarity of two learning resources  $i$  and  $j$  as shown in eq. (7).

$$sim(i, j) = \frac{\sum (R_{x,i} - \bar{R}_x)(R_{x,j} - \bar{R}_x)}{\sqrt{\sum (R_{x,i} - \bar{R}_x)^2} \sqrt{\sum (R_{x,j} - \bar{R}_x)^2}} \quad (7)$$

In equation (7),  $R_{x,i}$  denotes the score rated by the user  $x$  for the resource  $i$ ,  $R_{x,j}$  denotes the score rated by the user  $x$  for the resource  $j$ , and  $\bar{R}$  is the total mean value of the ratings given to the resource by multiple users. From the user-resource rating matrix in Fig. 4(a), a similarity matrix between different resources can be derived, as shown in

Fig. 4(b). The additional  $k$  ( $k \leq n$ ) resources that have the highest possible similarity to the current resource are selected to form an ensemble of resources that are highly similar to the current resource. From equation (8), the possible values of the preferred resource  $i$  for each user  $x$  can be predicted.

$$P_{x,i} = \frac{\sum_{t \in N} (sim(i, t) * R_{x,t})}{\sum_{t \in N} (\|sim(i, t)\|)} \quad (8)$$

In equation (8),  $N$  represents a set of resources similar to the resource  $i$  and  $R_{x,t}$  is the rating given to the resource by its user  $x$ . The predicted result matrix is shown in Fig. 5(a), while Fig. 5(b) shows the set of resource suggestions that are likely to be preferred for the user.

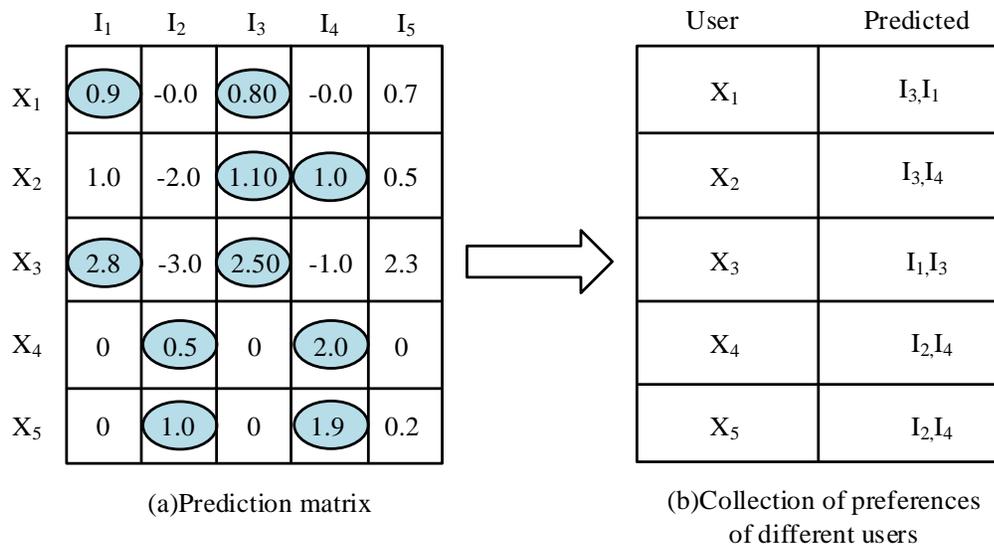


Fig.5. User prediction preference matrix.

The set of user-preferred resources generated by the collaborative filtering algorithm is obtained by similarity, and the order of these resource recommendations does not follow a specific learning path. Since the process of learning English is coherent, the list of recommended resources displayed in the learning platform should follow the learning path of the English subject. The study introduces a sequential pattern mining algorithm based on a collaborative filtering algorithm to arrange the list of English learning resources according to the learning order. In the web-based teaching platform, the key to the sequential pattern mining algorithm is to search for a series of resources that meet the least supported order of learning frequency on a specific set of input data sequences  $S$ , each list is a sequence of frequent learning resources, the specific algorithm flow is shown in Fig. 6.

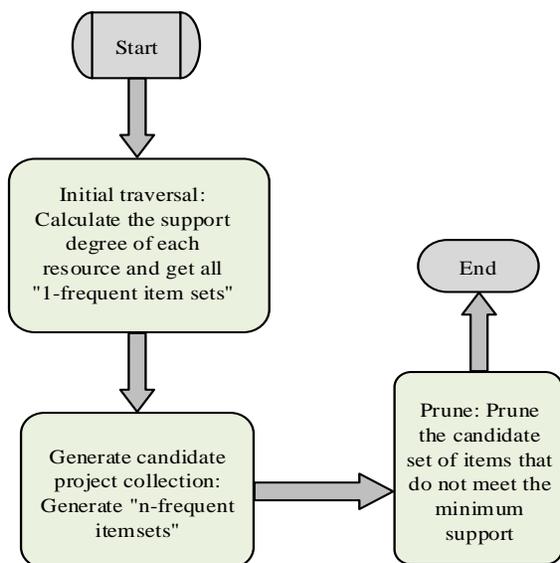


Fig.6. Flow chart of sequential pattern mining algorithm.

As can be seen from Fig. 6, the support degree of each resource is calculated during the initial search, and the set of all 1-frequency elements is obtained after the initial search, and the candidate set of n-frequency elements is generated from the set of n-1 ( $n > 1$ ) frequency elements. Finally, the candidate sets that do not meet the minimum support are removed by pruning branches. Since users may change their interest in learning

content over time when they are learning online, the user resource scores in the web logs do not objectively reflect the current user ratings and interests in the learning resources. The study introduces a damping function in the collaborative filtering algorithm to adjust the weight of learning resources rated by previous users, as shown in eq. (9).

$$Weight = \frac{2}{1 + e^{-(t-t_0)}} \quad (9)$$

In equation (9),  $t_0$  denotes the current time, and the closer the learning resource is to the current time, the higher weight is given to it, and the process of calculating the similarity between resources is also given a higher weight, when  $t \rightarrow t_0$ ,  $Weight \rightarrow 1$  and  $Value \rightarrow 1$ . Since the calculation of similarity between resources is performed when the user is offline, when calculating the relevance of n resources,  $O(n^2)$  space complexity is consumed to store the calculation results, increasing the space load and the efficiency of the recommendation algorithm decreases as the number of user visits increases. Therefore, the study selects only the most similar resources among the similar resources k as elements of the similar set in the computation process, which can greatly reduce the time complexity of the algorithm to a linear order of magnitude. In summary, the design of the study's proposed weblog-based online learning resource recommendation framework is shown in Fig. 7.

As can be seen in Fig. 7, offline pre-processing and filtering of user learning traces and resource records is performed to collect material that users may frequently access or want, and finally the merged resources are categorized according to the specific relationships between the different data to provide personalized services to users. And a tracking module is added to track whether users accept personalized recommendations from the platform as a way of tuning the data for future recommendations. To address the problem of new users with no learning history in online learning, the study groups and classifies new users based on the demographic information of users in the platform, and likewise groups and classifies new resources based on the resource topics in the resource database, thus solving the problem of cold processing in the design of the recommendation framework.

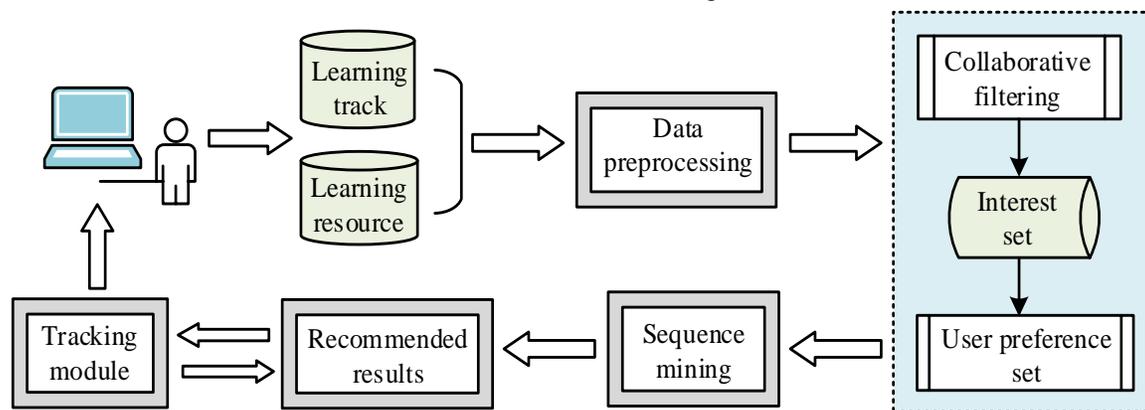


Fig.7. Web log-based recommendation framework for online learning resources.

#### IV SIMULATION TESTING OF THE ENGLISH SHARED RESOURCE PLATFORM AND THE PERFORMANCE OF PERSONALISED RECOMMENDATIONS

In order to demonstrate the reliable performance of the researched and developed universal platform for shared resources (denoted as Platform 1), the collaborative recommendation-based English teaching resource management platform proposed by Hu Ting scholars (denoted as Platform 2) and the blockchain-based virtual English teaching resource platform jointly designed by WANG P and QIAO S H.E (denoted as Platform 3) were used as experimental control, and the three platforms were used as a common platform for real English multimedia teaching materials on virtual hardware of the same scale and hardware standard. To make the experimental results more intuitive and to better prevent the problem of low variability of performance data due to the low concurrency of the system, the study selected 1200 sets of concurrent data with similar configuration parameters for testing. In addition, each test lasted around 3 minutes to fully account for the impact of the virtual environment and to protect the web hosts from high concurrent loads. The resource upload throughput results for each platform are shown in Fig. 8.

As can be seen from Fig. 8, as the concurrency continues to increase, the number of resource interactions handled by Platform 1 is higher than the other two benchmark platforms, and the curve changes less than the other two platforms. This indicates that the platform proposed in the study has a good load balancing capability in handling resource sharing requests, with each node of the platform handling user requests, which distributes the concurrency evenly to each node, allowing more resource requests to be handled. A graph of the response time results for each platform is shown in Fig. 9.

As can be seen in Fig. 9, the response times of the platforms created by the Institute are all lower than the response times of the benchmark platforms. Due to network congestion caused by concurrent transmissions, the response times of the platforms created by Platform 2 and Platform 3 increased dramatically, while failing to meet the requirements of a wide range of users, thus extending the response time of the shared resources. The results show that the platform designed in this study performs better, has better response speed and processing speed, can respond to users' daily access needs, and has strong load processing capacity.

To test the superior performance of the personalized recommendation model designed in this study, the experimental design and recommendations were compared using a weighted evaluation criterion, and the experimental results with different parameter settings were analyzed. Before evaluating the performance of the recommendation framework, the effect of the size of the parameter similarity set and the variation of the minimum support on the experimental results were determined. The study recommends resources to users based on similarity sets of different sizes and calculates the corresponding REM values for different minimum support cases, and the experimental results are shown in Fig. 10.

As can be seen from Fig. 10, the REM value decreases as the similarity set increases. However, the curve does not change significantly when the set of similarity goes from 7 to

10. Considering that the spatial complexity increases as the set of similarity increases, the study chose a similarity set size of 6 as the optimal value for subsequent experiments. The REM value changes as the minimum support increases. The REM value decreases when the minimum support is 0.1-0.5 and increases from 0.15 onwards. Therefore, in order to obtain more accurate recommendations, the minimum support was set to 0.15 in subsequent experiments. After the optimal parameter values were obtained, the parameters of each algorithm were set to the best value chosen and then tested. The results of the resource recommendations were categorized into three cases. The results of the collaborative filtering algorithm, the results of the generic sequence extraction algorithm, and the results obtained by a hybrid algorithm of collaborative filtering + sequence mining consisting of weighting, switching and merging, noting the above algorithms as Algorithm 1, Algorithm 2 and Algorithm 3, respectively. In the experiments, users pre-selected and ranked the topics of interest and their browsing. Their browsing history was also recorded by the platform. The resources of interest recommended by the platform to the users were then processed and calculated by different types of hybrid models and the correct recommendations were compared with the actual recommendations using the REM formula. The comparison of the recommendation results of each algorithm is shown in Fig. 11 and 12.

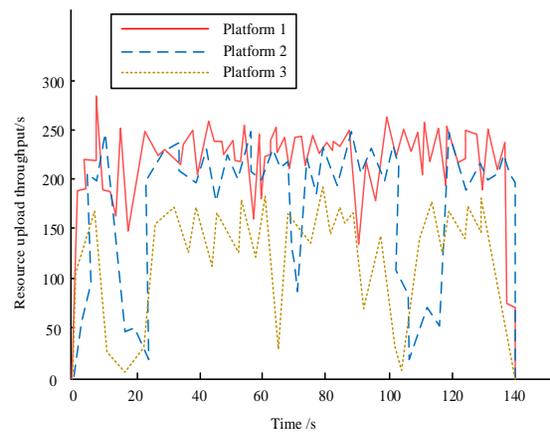


Fig.8. The upload throughput of English teaching resources on each platform.

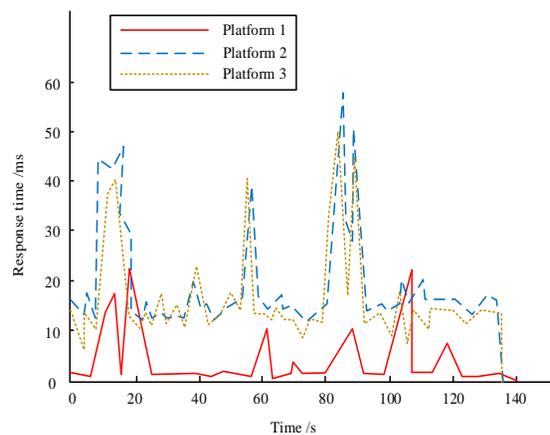


Fig.9. Response time by platform.

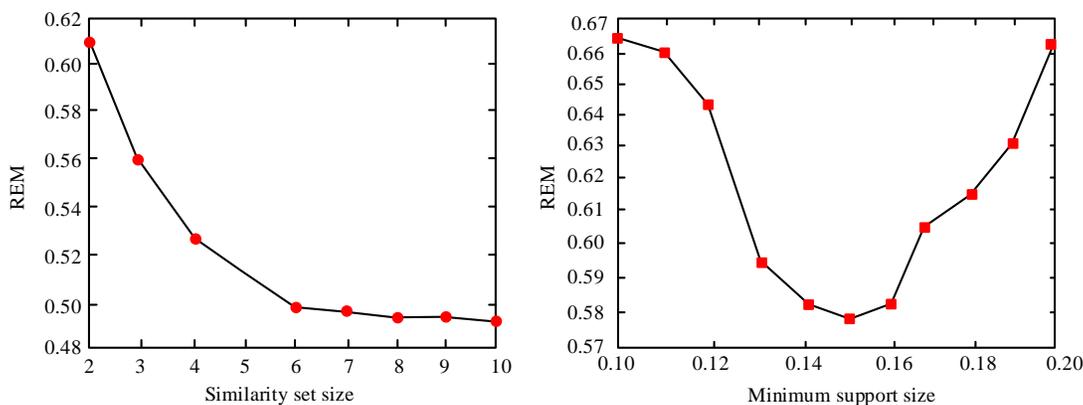


Fig.10. REM value variation under different similar set sizes and minimum support.

When the minimum support was chosen to be 0.15, the trend of REM changes as the size of the similar set expands from 2 to 10 as shown in Fig. 11. The simulation results show that as the similarity set expands, the REM values of all types of algorithms decrease, with the most obvious decreasing trend for the optimization algorithm 3 proposed in the study, and the REM value of algorithm 1 has almost no relationship with the size of the similarity set. The REM value of this hybrid algorithm model tends to be the smallest when the similarity set is greater than 6, which indicates that the algorithm is most accurate at a minimum support of 0.15, and the accuracy stabilizes when the similarity set size reaches 7.

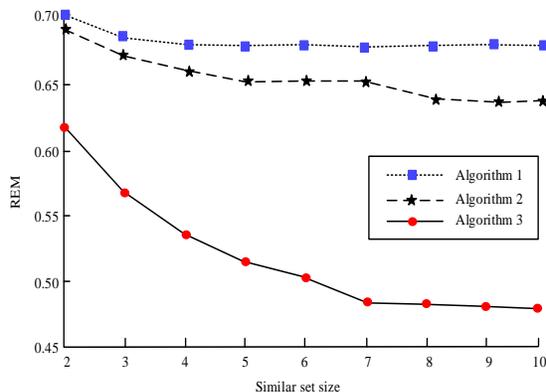


Fig.11. REM changes with the size of similar sets under specific support.

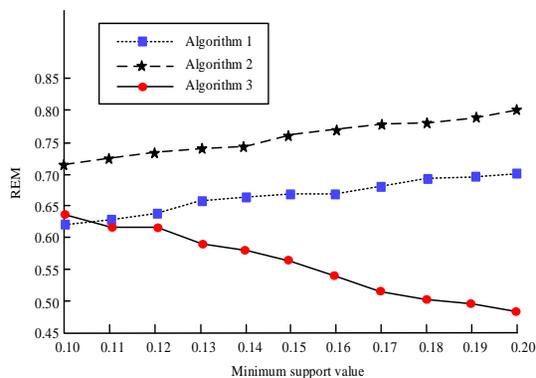


Fig.12. Change of REM with minimum support under specific similar set size.

When the size of the selected similarity set is 6, the trend of the REM values of each type of algorithm is shown in Fig. 12 as the minimum support degree increases from 0.1 to 0.2. The simulation results show that the larger the value of minimum support is taken, the gradually increasing REM value and decreasing accuracy of algorithms 1 and 2, while the continuous decreasing REM value and gradually increasing accuracy of the hybrid algorithm. In summary, the platform created by the research shows good performance in terms of resource transfer capability, achieving efficient teaching resource sharing in a short response time, and the proposed hybrid algorithm has the smallest REM value and the highest accuracy.

## V CONCLUSION

In response to the problems of scattered English online teaching resources that are difficult to achieve centralization and inefficient education resource recommendation, the study constructed an English online education resource sharing platform based on Web services. Simulation results show that the created platform performs well in terms of resource transfer capability and can achieve efficient teaching resource sharing in a short response time. The hybrid algorithm of collaborative filtering combined with sequence mining proposed in the study performs well in the validation of the recommendation algorithm. Compared with other algorithms, it has the smallest REM value and the highest accuracy rate. The simulation results prove that the English web-based educational resource sharing platform constructed by the study can perform the educational resource sharing task better and complete the English educational resource recommendation. There are also some shortcomings in this study; not enough user security authentication information is collected, and it is not effectively verified in the security module. It is expected that the in-depth research on this part will be strengthened in future research. In addition, the platform is not rich enough in features, and it is expected that more practical features will be added to the platform in the subsequent research to enhance the usefulness of the platform.

## REFERENCES

- [1] H. Nie, "Design and implementation of network mobile learning and hybrid learning platform based on MOOC platform", *IPPTA: Quarterly Journal of Indian Pulp and Paper Technical Association*, vol. 30(8), pp. 397-401, 2018.

- [2] Y. Sun, "An improved design method of English teaching system based on Multimedia Technology", *Modern electronic technology*, vol. 41(10), pp. 129-132, 2018.
- [3] T. Hu, "Design of English teaching resource management system based on collaborative recommendation", *Automation technology and application*, vol. 38(9), pp. 158-161, 2019.
- [4] H. Gao, L. Kuang, Y. Yin, et al. "Mining consuming Behaviors with Temporal Evolution for Personalized Recommendation in Mobile Marketing Apps", *Mobile Networks and Applications*, vol. 25(4), pp. 1233-1248, 2020.
- [5] S. Gu, X. Li, "Optimization of Computer-Aided English Translation Teaching Based on Network Teaching Platform", *Computer-Aided Design and Applications*, vol. 19(S1), pp. 151-160, 2021.
- [6] C. W. Yoo, H. C. Kim, "Dimensionality Reduction Method without Model for Personalized Recommendation", *Journal of Digital Contents Society*, vol. 21(3), pp. 587-592, 2020.
- [7] Q. Yuan, "Network education recommendation and teaching resource sharing based on improved neural network", *Journal of Intelligent and Fuzzy Systems*, vol. 39(4), pp. 5511-5520, 2020.
- [8] M. Gu, "Recursive Algorithm and its Practice in C Language Online Course Teaching", *Advances in Science and Technology*, vol. 105, pp. 341-347, 2021.
- [9] Y. Zhang, "Design and curriculum optimization of college english teaching model based on esp", *International Journal for Engineering Modelling*, vol. 31(1), pp. 359-364, 2018.
- [10] C. Fang, "Intelligent online teaching system based on SVM algorithm and complex network", *Journal of Intelligent and Fuzzy Systems*, vol. 40(5), pp. 1-11, 2020.
- [11] A. M. Christianson, "Using Socratic Online Polls for Active Learning in the Remote Classroom[J]", *Journal of Chemical Education*, vol. 97(9), pp. 2701-2705, 2020.
- [12] N. Zhou, Z. Zhang, J. Li, "Analysis on Course Scores of Learners of Online Teaching Platforms Based on Data Mining", *Ingénierie des Systèmes D'Information*, vol. 25(5), pp. 609-617, 2020.
- [13] Q. Zhang, Y. Liu, L. Liu, et al. "Location Identification and Personalized Recommendation of Tourist Attractions Based on Image Processing", *Traitement du Signal*, vol. 38(1), pp. 197-205, 2021.
- [14] S. E. Park, J. H. Yoon, S. Y. Choi, et al. "User-friendly Korean conversation learning application using face swap and personalized recommendation", *Journal of Digital Contents Society*, vol. 21(12), pp. 2125-2133, 2020.
- [15] Y. Xu, "Computer-Aided Design of Personalized Recommendation in Teaching System", *Computer-Aided Design and Applications*, vol. 17(S1), pp. 44-56, 2019.
- [16] H. Li, S. Zhang, J. Shi, et al. "Research and design of intelligent learning system based on recommendation technology", *Control & Intelligent Systems*, vol. 47(1), pp. 43-49, 2019.
- [17] Y. Chaabi, N. M. Ndiyaie, K. Lekdioui. "Personalized Recommendation Of Educational Resources In A MOOC Using A Combination Of Collaborative Filtering And Semantic Content Analysis", *International Journal of Scientific & Technology Research*, vol. 9(2), pp. 3243-3248, 2020.
- [18] K. Ohtomo, R. Harakawa, T. Ogawa, et al. "Personalized Recommendation of Tumblr Posts Using Graph Convolutional Networks with Preference-aware Multimodal Features", *ITE Transactions on Media Technology and Applications*, vol. 9(1), pp. 54-61, 2021.
- [19] X. Chen, Y. Xue, Y. Shiue, "Rule based Semantic Reasoning for Personalized Recommendation in Indoor O2O e-commerce", *International Core Journal of Engineering*, vol. 6(1), pp. 309-318, 2020.
- [20] X. Wang, "Personalized recommendation framework design for online tourism: know you better than yourself", *Industrial Management & Data Systems*, vol. 120(11), pp. 2067-2079, 2020.
- [21] M. M. Juraev, "Prospects for the development of professional training of students of professional educational institutions using electronic educational resources in the environment of digital transformation", *Academia Globe: Inderscience Research*, vol. 3(10), pp. 158-162, 2022.
- [22] A. K. Barianos, A. Papadakis, N. Vidakis, "Content manager for serious games: Theoretical framework and digital platform", *Advances in Mobile Learning Educational Research*, vol. 2(1), pp. 251-262, 2022.

# A Study of Encryption for Multimedia Digital Audio Security

Xiaodong Zhou\*, Chao Wei, Xiaotang Shao

School of Literature-Journalism and Communication, Sanjiang University, Nanjing, Jiangsu 210012, China

**Abstract**—Driven by the development of multimedia, the encryption of multimedia digital audio has received more attention; however, cryptography-based encryption methods have many shortcomings in encryption of multimedia information, and new encryption methods are urgently needed. This paper briefly introduced cryptography and chaos theory, designed a chaos-based encryption algorithm that combined Logistic mapping and Sine mapping for confusion and used a Hopfield chaos neural network for diffusion, explained the encryption and decryption process of the algorithm, and tested the algorithm. It was found that the keys obtained by the proposed algorithm passed the SP800-22 test, and the correlation between the three encrypted audio and the original audio was 0.0261, -0.0536, and 0.0237, respectively, all of which were small, and the peak signal-to-noise ratio (PSNR) values were -0.348 dB, -7.645 dB, and -3.636 dB, respectively, which were significantly different from the original audio. The NSCR and UACI were also closer to the original values. The results prove that the proposed algorithm has good security and can encrypt the actual multimedia digital audio.

**Keywords**—Multimedia digital audio; chaotic theory; encryption; logistic mapping; sine mapping; security

## I. INTRODUCTION

The dissemination speed of multimedia information is increasingly accelerated with the development of computer technology [1]. Relying on the Internet, mobile terminals, etc., digital images, video, audio, and other multimedia information is generated and transmitted all the time, which facilitates people's communication and exchange and also brings new challenges to information security. Multimedia information is mostly transmitted and stored in public environment, and under the influence of network, it spreads faster and wider, and the danger of information leakage is also greater [2]. Encryption can effectively improve the security of multimedia information, so multimedia encryption has also become an important content [3]. At present, many methods have been applied in the encryption of texts and images; however, compared with them, audio has greater redundancy and higher relevance, so the traditional text and image encryption methods are not applicable to audio. Therefore, encryption for digital audio has become a common concern for researchers [4]. Singh et al. [5] compared the performance of dynamic advanced encryption standard (AES) and standard AES for audio encryption and analyzed the quality of the algorithms by histogram, correlation, etc. Babu et al. [6] converted audio data to image data, studied the encryption and decryption of audio using a fractional order hyperchaotic system, and verified the security of the system through analysis. Wang et al. [7] proposed an encryption

method using a chaotic system and deoxyribonucleic acid (DNA) coding and found that the method performed well in multichannel audio processing through comparative experiments on different types of audio. Zaid et al. [8] proposed two chaos-based permutation algorithms: Arnold cat mapping and Baker mapping. The experiments showed that both algorithms can provide reliable security, but in most cases, Arnold cat mapping performs better. At present, there are still many challenges in multimedia digital audio encryption, and the security of existing methods cannot meet such encryption needs yet. Therefore, in order to find out a more suitable encryption method for multimedia digital audio, this paper designed a chaos-based method and proved the reliability of the method through experiments. This work provides a new method for the research of multimedia digital audio encryption and also provides theoretical support for the in-depth research of multimedia information encryption. This paper first briefly introduces cryptography and chaos theory in Section II. It describes the encryption and decryption method based on Logistic mapping, Sine mapping, and Hopfield chaotic neural network designed in this paper. Section III presents the experiments on the proposed encryption and decryption method used to prove its security for multimedia digital audio encryption and decryption. Section IV is the conclusion section, which briefly summarizes and reflects on the research of this paper.

## II. CHAOS-BASED AUDIO ENCRYPTION ALGORITHM

### A. Cryptography and Chaos Theory

A simple password system generally consists of several components, as shown in Fig. 1.

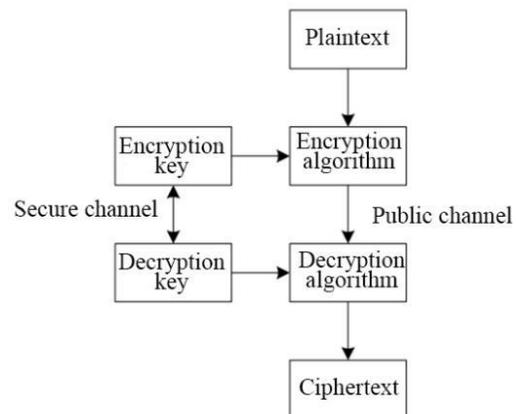


Fig. 1. A simple password system.

As shown in Fig. 1, the plaintext is the original message to be encrypted, written as  $M$ . The ciphertext is the encrypted message, written as  $C$ . It is assumed that there is an encryption algorithm  $E$ , then the encryption process is written as:  $E(M) = C$ . Let the decryption algorithm be  $D$ , then the decryption process is written as:  $D(C) = M$ .

For audio information with high redundancy and high correlation, traditional encryption algorithms, such as AES and DES [9], are unable to encrypt effectively. Chaos contains characteristics such as ergodic, unpredictable, and random, and it can be applied to encryption to get good results [10]. In the Devaney's definition of chaos [11], for mapping  $f$  in the metric space  $V$ , if it is chaotic, then the following conditions are satisfied:

- 1) There exists  $\delta > 0$ , for any  $\varepsilon > 0$  and  $x \in V$ , there exists  $y$  and natural number  $n$  in the  $\varepsilon$  neighbourhood of  $x$  such that  $d[f^n(x), f^n(y)] > \delta$ ;
- 2) For any open sets  $X$  and  $Y$  in  $V$ , there exists  $k > 0$  such that  $f^k(X) \cap Y \neq \emptyset$ ;
- 3) The periodic orbit of  $f$  is dense in  $V$ .

Chaos is usually determined using the Lyapunov exponential method [12]. In a one-dimensional chaotic system, there exists an orbit:  $x_0, x_1 = f(x_0), \dots, x_n = f(x_{n-1}), \dots$ . A perturbation  $\delta x_0$  is added to  $x_0$ . After  $n$ -step iterations, the resulting perturbation is written as:

$$\delta x_n = f'(x_{n-1})f'(x_{n-2}) \dots f'(x_0)\delta x_0. \quad (1)$$

The Lyapunov exponent is written as:

$$\Lambda = \lim_{\delta x_0 \rightarrow 0} \lim_{n \rightarrow \infty} \frac{1}{n} \log \left| \frac{\delta x_n}{\delta x_0} \right| = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=0}^{n-1} \log |f'(x_i)|. \quad (2)$$

When  $\lambda > 0$ , it means that the orbit is sensitive to the initial value, i.e., it is a chaotic orbit.

Classical chaotic systems include the following types.

1) Logistic mapping [13]:  $x_{r+1} = \mu x_r(1 - x_r)$ , where  $r$  is the number of iterations and  $\mu$  is the system bifurcation parameter,  $\mu \in (0,4)$ . When  $3.5699456 < \mu \leq 4$ , the system is in a chaotic state.

2) Henon mapping [14]:  $\begin{cases} x_{r+1} = -px_r + y_r + 1 \\ y_{r+1} = qx_r \end{cases}$ . When  $p = 1.76$  and  $q = 0.1$ , the system is in a chaotic state.

3) Sine mapping [15]:  $x_{r+1} = \mu \sin(\pi x_r)$ . When  $\mu \in [3.48, 4]$ , the system is in a chaotic state.

4) Lorenz chaotic system [16]:  $\begin{cases} x' = -\delta(x - y) \\ y' = -xz + \gamma x - y \\ z' = \alpha x - bz \end{cases}$ , where  $\delta, \gamma$ , and  $b$  are the system control parameters. When  $\delta = 10, \gamma = 28, b = 8/3$ , or  $\delta = 16, \gamma = 40, b = 4$ , the system is in a chaotic state.

### B. Audio Encryption Algorithm

One-dimensional chaotic mapping is simple in the chaotic system. In order to improve the security of chaotic encryption, this paper proposes an improved method, i.e., combining two one-dimensional chaotic mappings. Logistic mapping has the problem of uneven data distribution, and the same defect also

exists in the Sine mapping. Therefore, they are combined to obtain the Logistic-Sine-coupling mapping (LSCM), and the corresponding equation is:

$$x_{r+1} = (\mu x_r(1 - x_r) + (4 - \mu) \sin(\pi x_r) / 4) \bmod 1. \quad (3)$$

When  $\mu \in (3, 4]$ , the system is in a chaotic state.

With the continuous development of neural networks, their applications in fields such as artificial intelligence are becoming more and more widespread, and neural networks also carry the chaotic characteristics. Hopfield neural networks are enough to meet the requirements of cryptography and have good performance in encryption [17]. It is divided into two types, discrete and continuous. The discrete type is used in this paper, and its expression is:

$$x = -x_i + \sum_{i=1}^3 w_{ij} v_i, \quad (4)$$

$$v_i = \tanh(x_i) = \frac{e^{x_i} - e^{-x_i}}{e^{x_i} + e^{-x_i}}, \quad (5)$$

where  $w_{ij}$  is the weight matrix. The three-dimensional Hopfield neural network with high operational efficiency and a good chaotic state is called Hopfield chaotic neural network (HCNN), and the corresponding equation is:

$$\begin{cases} x'_1 = -x_1 + 2 \tanh(x_1) - \tanh(x_2) \\ x'_2 = -x_2 + 1.7 \tanh(x_1) + 1.7 \tanh(x_2) + 1.1 \tanh(x_3) \\ x'_3 = -x_3 - 2.5 \tanh(x_1) - 2.9 \tanh(x_2) + 0.56 \tanh(x_3) \end{cases} \quad (6)$$

In multimedia digital audio encryption and decryption, the LS mapping is used to perform confusion operation on audio, and then HCNN is used to generate diffusion sequence. First, the encryption process is as follows:

- 1) The original audios from the left and right channels are read and denoted as two sets of audio  $A (L \times 2)$ .
- 2) Hash operation is performed on the original audios to get hashed value  $h$ :  $h = \text{hash}(A, 'SHA - 512')$ .
- 3) The key generation process is as follows.  $\text{hex2dec}$  is a function that converts a hexadecimal hash code to a decimal number, and  $m$  is the number of iterations.

$$\begin{cases} x_{01} = \text{hex2dec}(h(1:25)) / (L \times 10^{24}) \\ x_{02} = \text{hex2dec}(h(26:50)) / (L \times 10^{24}) \\ x_{03} = \text{hex2dec}(h(51:75)) / (L \times 10^{24}) \\ x_{04} = \text{hex2dec}(h(76:100)) / (L \times 10^{24}) \\ x_{05} = \text{hex2dec}(h(101:125)) / (L \times 10^{24}) \\ m = 10000 + \text{hex2dec}(h(126:128)) \end{cases} \quad (7)$$

4) Initial values  $x_{01}$  and  $x_{02}$  are processed to obtain initial values  $x_1$  and  $\mu$  of LSCM:  $\begin{cases} x_1 = \text{mod}(x_{01}, 1) \\ \mu = \text{mod}(x_{02}, 1) \end{cases}$ , where  $\text{mod}$  is the modulo operation. Then, the LSCM is subjected to  $m$  preiteration to fully reach the chaotic state, and then it is iterated  $2L$  times to obtain the chaotic sequence:  $\begin{cases} X_1 = \{x_{11}, x_{12}, \dots, x_{1L}\} \\ X_2 = \{x_{21}, x_{22}, \dots, x_{2L}\} \end{cases}$

5) Random sequences  $X - H_1$  and  $X - H_2$  without repetition are generated based on  $X_1, X_2$ , and two arrays of natural numbers  $H_1 = \{1, 2, \dots, L\}$  and  $H_2 = \{1, 2, \dots, L\}$  to confuse audio A. Then,  $A'$ :  $\begin{cases} A'(1:L, 1) = A(X - H_1(1:L), 1) \\ A'(1:L, 2) = A(X - H_2(1:L), 2) \end{cases}$  is obtained.

6)  $x_{03}, x_{04}$ , and  $x_{05}$  are substituted into the three-dimensional HCNN to obtain three diffusion sequences:  $\begin{cases} Y_1 = \{y_{21}, y_{22}, \dots, y_{2L}\} \\ Y_2 = \{y_{31}, y_{32}, \dots, y_{3L}\} \\ Y_3 = \{y_{41}, y_{42}, \dots, y_{4L}\} \end{cases}$ .

7) Exclusive OR diffusion is performed to obtain encrypted speech  $C$ :  $\begin{cases} C_1(i) = bitXOR(B_1(i-1), Y_1(i)) \\ C_2(i) = bitXOR(B_2(i-1), Y_2(i)) \end{cases}$ , where  $bitXOR$  is the bitwise exclusive OR function and  $Y_1, Y_2$  and  $Y_3$  are the chaotic sequence obtained by HCNN.

8) To further improve the encryption performance,  $Y_1, Y_2$ , and  $Y_3$  are combined two by two for three times of diffusion to obtain the final encrypted speech and complete the encryption of the audio.

The decryption process of multimedia digital audio is as follows:

- 1) The encrypted audio is read.
- 2) Initial values are obtained using LSCM and HCNN in accordance with the same steps as encryption to get chaotic sequences  $X_1$  and  $X_2$  needed for decryption.
- 3)  $X_1$  and  $X_2$  are used to obtain decrypted diffusion sequences  $Y_1, Y_2$ , and  $Y_3$ .
- 4) The encryption process is reversed to perform decryption diffusion on the encrypted audio, followed by confusion. Finally, the decrypted audio is obtained.

### III. AUDIO ENCRYPTION ALGORITHM SECURITY ANALYSIS

Experiment was carried out in Windows 10 environment, 3.4GHz processor, and 4G RAM. In the chaotic system, the value of  $\mu$  was set as 3.707 and 3.808, respectively, and initial values  $x_{01} = 0.7, x_{02} = 0.8$ . The audios to be tested were all in wave format. The first three audios, named audio1.wav, audio2.wav, and audio3.wav, came from the Internet, and the other three audios came from THCHS-30 voice library [18]. Audios in THCHS-30 voice library were collected in a quite office environment at a sampling frequency of 16 kHz, the total duration of those audios was 30 hours, and the sampling size was 16 bits. Three audios were randomly selected from the library for experiments, named audio4.wav, audio5.wav, and audio6.wav. Taking audio1.wav as an example, the result of encryption and decryption using the proposed method is shown in Fig. 2.

Fig. 2 shows the original audio waveform of audio1.wav, and Fig. 3 shows the audio waveform obtained after audio1.wav was encrypted. It was found from the comparison between Fig. 2 and 3 that the encrypted audio did not have similarities with the original audio and was not associated with the original audio, which showed that the audio encryption method was effective and could encrypt the audio well. Fig. 4 shows the audio waveform obtained after decrypting the encrypted audio. The comparison between Fig. 2 and 4 showed that the correct original audio was obtained after decrypting

using the proposed method, which proved the usability of the method.

First, the randomness of the key was tested using 15 items in the SP800-22 test package from National Institute of Standards and Technology (NIST) test, and the randomness was judged by the P value. The higher the P value, the stronger is the randomness. The results of the key test are displayed in Table I.

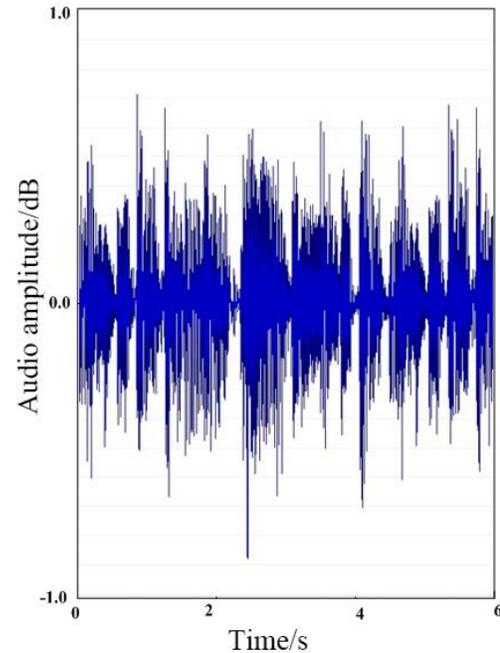


Fig. 2. Original audio.

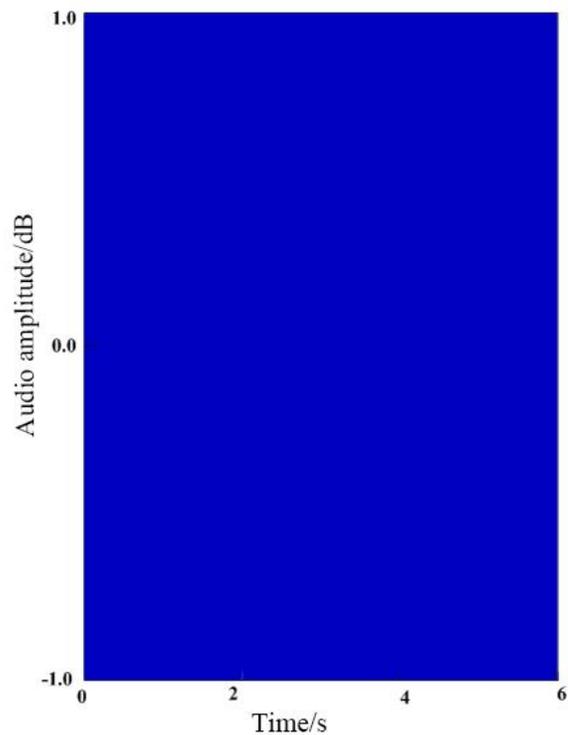


Fig. 3. Encrypted audio.

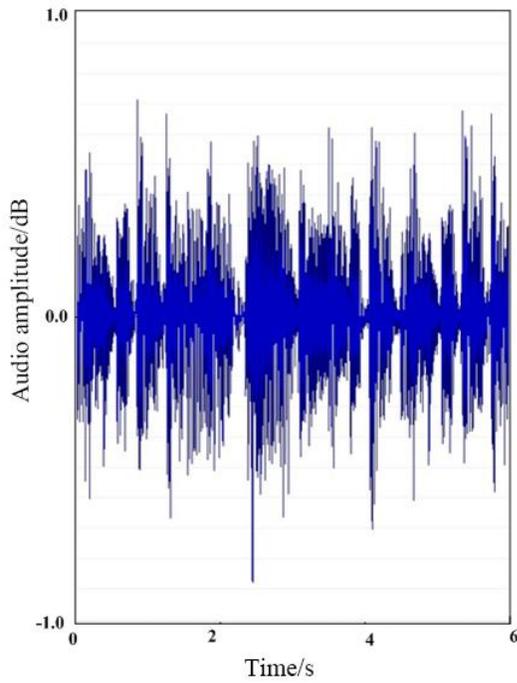


Fig. 4. Decrypted audio.

TABLE I. SP800-22 TEST RESULTS

Statistical test	P value	Result
Frequency	0.6788	Pass
Block Frequency	0.4795	Pass
Cumulative Sums	0.0945	Pass
Runs	0.3152	Pass
Longest Run	0.0528	Pass
Rank	0.7958	Pass
FFT	0.5746	Pass
Non Overlapping Template	0.9954	Pass
Overlapping Template	0.9925	Pass
Universal	0.1452	Pass
Approximate Entropy	0.9258	Pass
Random Excursions	0.6521	Pass
Random Excursions Variant	0.9654	Pass
Serial	0.4215	Pass
Linear Complexity	0.8752	Pass

It was seen from Table I that the keys generated using the proposed method could pass the SP800-22 test, and the P values were all greater than 0.01, indicating that the keys had good randomness and were suitable for encrypting multimedia digital audio.

The correlation coefficient reflects the correlation between two data. If there is a small correlation coefficient between the encrypted audio and the plaintext audio, it means the less similarity between the plaintext and the ciphertext. The correlation coefficient is calculated as follows:

$$r = \frac{\sum_{i=1}^n (A(i) - \bar{A})(B(i) - \bar{B})}{\sqrt{\sum_{i=1}^n (A(i) - \bar{A})^2 \sum_{i=1}^n (B(i) - \bar{B})^2}} \quad (8)$$

where  $\bar{A}$  and  $\bar{B}$  are the mean values of A and B. The correlation coefficient of the audio before and after the encryption by the proposed method was calculated, and the results were compared with Mohamed's method [19], as shown in Fig. 5.

It was observed in Fig. 5 that the correlation between the six encrypted test audios and the original audio was small, and the coefficients were 0.0261, -0.0536, 0.0237, 0.0227, -0.0577, and 0.0219, respectively. Compared with Mohamed's method [19], the audio correlation before and after encryption by the method proposed in this paper was smaller, indicating that the similarity between the ciphertext and the plaintext was lower, i.e., the method was safe.

The peak signal-to-noise ratio (PSNR) reflects the quality of signal compression. The larger the value of PSNR, the better is the quality of signal compression, and the closer to the original audio. Conversely, if the PSNR value of the encrypted audio is smaller, it means that it is more different from the original audio. The PSNR calculation formula is:

$$PSNR = 10 \log_{10} \left( \frac{\max A}{\sqrt{MSE}} \right)^2 \quad (9)$$

$$MSE = \frac{1}{M \times N} \sum_{i,j} (A(i,j) - B(i,j))^2 \quad (10)$$

where M and N are the width and height of the audio, A and B are the original and encrypted audio. The PSNR obtained by the method proposed in this paper was compared with the results in Tamimi's study [20] and Liu's study [21], as shown in Fig. 6.

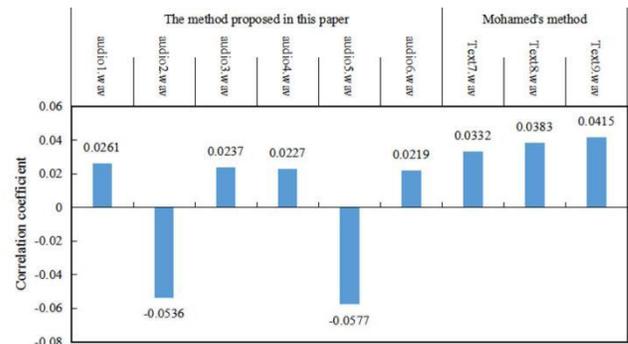


Fig. 5. Comparison of correlation coefficients before and after encryption.

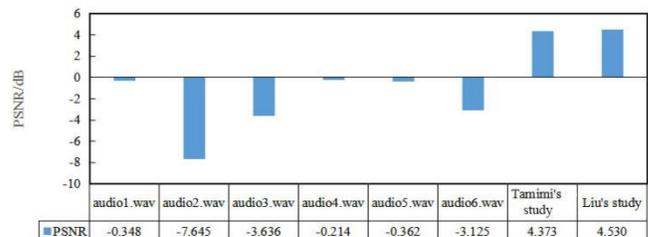


Fig. 6. Comparison of PSNR.

It was observed in Fig. 6 that the PSNR of the six audios were -0.348 dB, -7.645 dB, and -3.636 dB, which were small, and the PSNR was 4.373 dB in Tamimi's study [20] and 4.530 dB in Liu's study [21]. The PSNR values obtained in this paper were smaller; indicating that the audios encrypted by the method proposed in this paper had higher security and was more resistant to attacks.

Finally, the performance of this method against differential attacks was analyzed based on the indexes of the number of samples changes rate (NSCR) and the uniform average change intensity (UACI). The following equations are:

$$NSCR = \frac{\sum_{i=1}^L |Sign(B(i)-B'(i))|}{L} \times 100\%, \quad (11)$$

$$UACI = \frac{1}{L} \sum_{i=1}^L \frac{|B(i)-B'(i)|}{2^{8-1}} \times 100\%, \quad (12)$$

where  $B(i)$  is the encrypted audio,  $B'(i)$  is the encrypted audio with one original audio sampling data randomly changed, and  $Sign$  is the sign function. When the audio signal was 8 bit, the ideal values of NSCR and UACI were 100% and 33.33%, respectively. The average values were taken after several tests and compared with the results in Soliman's study [22] and Shah's study [23], and the results are shown in Fig. 7.

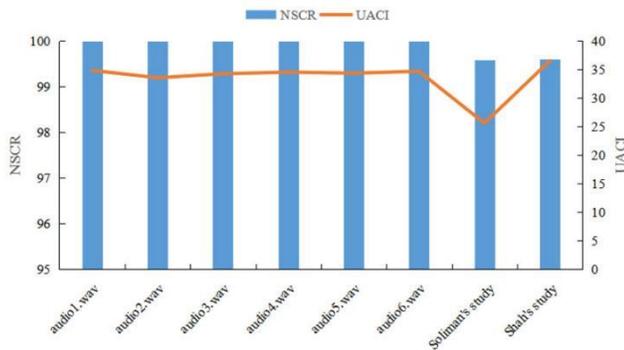


Fig. 7. Comparison of NSCR and UACI.

It was observed in Fig. 7 that compared with Soliman's study [22] and Shah's study [23], the NSCR obtained by the proposed method was always above 99.99%, which was closer to the ideal value (100%), and the UACI obtained by the proposed method was 34.8542%, 33.5628%, 34.2587%, 34.5515%, 34.3637%, and 34.6987%, which was closer to the ideal value (33.33%). These results verified the performance of the chaos-based audio encryption method in resisting differential attacks.

It was concluded from the above experimental results that the method proposed in this paper had a good encryption and decryption performance for multimedia digital audio, the encrypted audio files were not similar to the original files, and the original audio was well recovered after decryption. From the security point of view, the key obtained by the method had good randomness and passed the SP800-22 test. Then, from the comparison of different indicators, the experiments on six different audios revealed that the correlation between the audio before encryption and after decryption obtained by the method was very small, and the PSNR was also significantly smaller compared with the results in other literature, suggesting good

resistance to attacks. The experimental results prove the superiority of the method for multimedia digital audio encryption and the reliability of the encryption method combining different chaos methods and further verify the usability of chaos theory for multimedia information encryption.

#### IV. CONCLUSION

This paper designed a chaos-based encryption method for the encryption of multimedia digital audio, combined LSCM with HCNN to realize the encryption of digital audio, and analyzed its security. It was found that the key obtained by the proposed method could pass the SP800-22 test, with good randomness, and the encrypted audio had less correlation with the original audio (below 0.03), smaller PSNR value, above 99.99% NSCR value, its UACI was closer to the ideal value (33.33 %), and its resistance to differential attacks was strong. The method can be further applied in practical multimedia digital audio encryption. However, there are also some shortcomings in this paper, such as the small scale of experimental data and no practical application. In future research, further studies can be conducted in hardware implementation and encryption system design to understand the operability of the method in a practical environment.

#### REFERENCES

- [1] H. Aziz, S. Gilani, I. Hussain, A. K. Janjua, and S. Khurram, "A Noise-Tolerant Audio Encryption Framework Designed by the Application of S 8 Symmetric Group and Chaotic Systems," *Math. Probl. Eng.*, vol. 2021, pp. 5554707.1-5554707.15, April 2021.
- [2] N. Sen, R. Dantu, and M. Thompson, "Performance Analysis of Elliptic Curves for VoIP Audio Encryption Using a Softphone," *International Conference on Security and Privacy in Communication Systems*, pp. 503-508, December 2020.
- [3] S. Eldin, S. A. Khamis, A. Hassanin, and M. A. Alsharqawy, "New audio encryption package for TV cloud computing," *Int. J. Speech Technol.*, vol. 18, pp. 131-142, March 2015.
- [4] L. Zhou, X. Li, F. Tan, Y. Huang, and W. Ma, "A two-layer networks-based audio encryption/decryption scheme via fixed-time cluster synchronization," *Soft Comput.*, vol. 26, pp. 9761-9774, July 2022.
- [5] A. Singh, P. Agarwal, and M. Chand, "A Comparative Study of Audio Encryption Analysis Using Dynamic AES and Standard AES Algorithms," *International Workshop Soft Computing Applications*, pp. 241-249, January 2021.
- [6] N. R. Babu, M. Kalpana, and P. Balasubramaniam, "A novel audio encryption approach via finite-time synchronization of fractional order hyperchaotic system," *Multimed. Tools Appl.*, vol. 80, pp. 1-25, February 2021.
- [7] X. Wang, and Y. Su, "An Audio Encryption Algorithm Based on DNA Coding and Chaotic System," *IEEE Access*, vol. 8, pp. 9260-9270, 2020.
- [8] O. Zaid, M. A. Tawfeek, and S. Alanazi, "Applying and Comparison of Chaotic-Based Permutation Algorithms for Audio Encryption," *Comput. Mater. Con.*, vol. 67, pp. 3161-3176, February 2021.
- [9] M. Karmani, N. Benhadjoussef, B. Hamdi, and M. Machhout, "The DFA/DFT-based hacking techniques and countermeasures: Case study of the 32-bit AES encryption crypto-core," *IET Comput. Digit. Tec.*, vol. 15, pp. 160-170, March 2021.
- [10] K. Mistry, S. Dash, and S. Tallur, "Audio encryption through synchronization of chaotic oscillator circuits: Teaching non-linear dynamics through simple electrical circuits," *Am. J. Phys.*, vol. 87, pp. 1004-1013, December 2019.
- [11] H. Wang, Q. Liu, H. Li, and H. Fu, "Sensitivity, Devaney's chaos and Li-Yorke  $\epsilon$ -chaos," *Semigroup Forum*, vol. 100, pp. 888-909, February 2020.

- [12] M. Shafiya, and G. Nagamani, "New finite-time passivity criteria for delayed fractional-order neural networks based on Lyapunov function approach," *Chaos Soliton. Fract.*, vol. 158, pp. 1-12, May 2022.
- [13] S. Kanwal, S. Inam, O. Cheikhrouhou, K. Mahnoor, A. Zaguia, and H. Hamam, "Analytic Study of a Novel Color Image Encryption Method Based on the Chaos System and Color Codes," *Complexity*, vol. 2021, pp. 5499538-1-5499538-19, June 2021.
- [14] X. Zhuo, M. Bi, Z. Hu, H. Li, X. Wang, and X. Yang, "Secure scheme for OFDM-PON system using TR based on modified Henon chaos," *Opt. Commun.*, vol. 462, pp. 1-7, May 2020.
- [15] Z. Pan, W. Lu, H. Wang, and Y. Bai, "Recognition of a linear source contamination based on a mixed-integer stacked chaos gate recurrent unit neural network-hybrid sparrow search algorithm," *Environ. Sci. Pollut. R.*, vol. 29, pp. 33528-33543, May 2022.
- [16] J. Shen, B. Liu, Y. Mao, R. Ullah, J. Ren, J. Zhao et al., "Enhancing the Reliability and Security of OFDM-PON Using Modified Lorenz Chaos Based on the Linear Properties of FFT," *J. Lightwave Technol.*, vol. 39, pp. 4294-4299, April 2021.
- [17] C. Lakshmi, K. Thenmozhi, J. B. B. Rayappan, and R. Amirtharajan, "Hopfield attractor-trusted neural network: an attack-resistant image encryption," *Neural Comput. Appl.*, vol. 32, pp. 11477-11489, August 2020.
- [18] D. Wang, and X. Zhang, "THCHS-30 : A Free Chinese Speech Corpus," arXiv e-prints, December 2015.
- [19] A. A. Mohamed, M. Ismail, and N. Zainal, "Robust Audio Encryption Method for MPEG-2 AAC Audio Based on Module Arithmetic and Chaotic Maps," *IRECOS*, vol. 10, pp. 80-89, January 2015.
- [20] A. A. Tamimi, and A. M. Abdalla, "An audio shuffle-encryption algorithm," *Lect. Notes Eng. Comput. Sci.*, vol. 2213, pp. 409-412, October 2014.
- [21] H. Liu, A. Kadir, and Y. Li, "Audio encryption scheme by confusion and diffusion based on multi-scroll chaotic system and one-time keys," *Optik*, vol. 127, pp. 7431-7438, May 2016.
- [22] N. F. Soliman, M. I. Khalil, A. D. Algarni, S. Ismail, R. Marzouk, and W. El-Shafai, "Efficient HEVC Steganography Approach Based on Audio Compression and Encryption in QFFT Domain for Secure Multimedia Communication," *Multimed. Tools Appl.*, vol. 80, pp. 4789-4823, January 2021.
- [23] D. Shah, T. Shah, and S. S. Jamal, "Digital audio signals encryption by Mobius transformation and Hénon map," *Multimedia Syst.*, vol. 26, pp. 235-245, April 2020.

# Digital Twins for Smart Home Gadget Threat Prediction using Deep Convolution Neural Network

Mrs. Valluri Padmapriya, Dr. Muktevi Srivenkatesh

Department of Computer Science-GITAM School of Science, GITAM Deemed to be University,  
Visakhapatnam, Andhra Pradesh, India

**Abstract**—Digital twin is one of the most important innovations in the Internet of Things (IoT) era and business disruption. Digital twins are a growing technology that bridges the gap between the real and the digital. Home automation in the IoT refers to the practice of automatically managing and monitoring smart home electronics by use of a variety of control system methods. The geysers, refrigerators, fans, lighting, fire alarms, kitchen timers, and other electrical and electronic items in the home can all be managed and monitored with the help of a variety of control methods. Digital twins replicate the physical machine in real time and produce data, such as asset degradation, product performance level that may be used by the predictive maintenance algorithm to identify the product functionality levels. The purpose of this research is to design the framework of Digital Twin using machine learning and state estimation algorithms model to assess and predict home appliances based on the probability rate of smart home system gadgets functionality. The main goal of this research is to create a digital twin for smart home gadgets that are used to monitor the health status of these devices for increasing the life time and to reduce maintenance costs. This research presents a Deep Convolution Neural Network based Logistic Regression Model with Digital Twins (DCNN-LR-DT) for accurate prediction of smart home gadget functionality levels and to predict the threats in advance. The proposed model is compared with the traditional models and the results represent that the proposed model performance is better than traditional models.

**Keywords**—Digital twins; deep learning; convolution neural network; logistic regression; internet of things; smart home; IoT gadget functionality; threat prediction

## I. INTRODUCTION

Modern industry and the country's economy depend critically on Industrial revolution 4.0 and smart manufacturing. Industry 4.0 intends to create a global networked infrastructure that addresses compatibility and interoperability problems within and between all levels of automated systems and factories, enhancing the agility and flexibility of manufacturing methods [1]. Advanced robotics, which acts as an agent acting that appears in every area of production lines, is equally essential to smart factory. Digital Twin (DT) is attracted increasing scientific attention as a result of the extensive research and development on Industry 4.0 and Artificial Intelligence (AI) [2]. The network and data serve as the foundation of DT as a digital representation of a physical

object, the algorithm and modeling serve as the core, and the application and service serve as the application. The expenses of manufacturing businesses rise as a result, and at the same time, their organizational structures and operational procedures face enormous difficulties [3]. In light of this, AI-powered DT technology is anticipated to adapt conventional model-based techniques to changing boundary conditions and offer a demand-oriented, real-time competent evaluation basis to effectively assist decision-making in multi-objective challenges [4]. Numerous studies have already discussed and characterized DT from the standpoint of broader concepts and technology, as well as some sectors, without a specific focus on AI, such as product design, modelling and simulation, and problem diagnosis and prognostics [5]. Various engineering implementation scenarios pose unique problems [6]. The systematic and thorough integration of domain-specific knowledge is much more crucial for the foundation of DT and AI. In the context of the environment and circular economy, there is currently still a dearth of thorough industry-focused reviews of AI and DT technology.

The National Aerospace and Space Administration describes the idea of DT as an act involving, multiscale, statistical simulation that makes use of physical models, sensor feeds, fleet histories, etc. to mimic the twin's daily activities [7]. Any corporation can gain from having electronic information, and even an individual may value their data to the point that they simply cannot risk losing it. In recent years, the availability of low-cost sensors and open-source middle-ware software has opened up interesting research areas in the robotics field. In particular, the next generation of simulation models called DT [8], which represents a continuous virtual replica of physical systems, has gained increasing attention. It can be used to create a simulation of a smart home including assistive/service robots and human users. It has applications in the optimization of robots and smart home settings [9]. For example, finding the optimal number and configuration of sensors especially when new robots or users are introduced [10]. As another example, monitoring in real-time, further analysis and learning of edge cases and rare situations in DT for the safety of human users, e.g., by pushing those events from simulation to real-world and vice-versa. In practice, such optimization should be carried out over a variety of houses and users. The digital twin can be created in all scenarios [11]. The digital model and process is shown in Fig. 1.

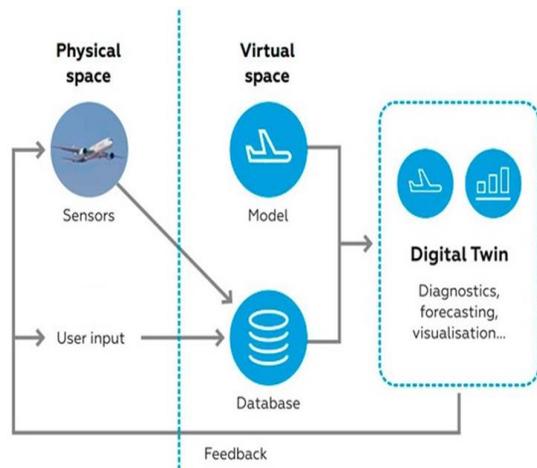


Fig. 1. Digital twin.

With the mounting deployments of the Internet of Things (IoT) systems, the significance of the concept of a digital illustration of physical things has gathered trivial interest in the recent years. Digital Twin is basically a living model of the physical skill or system, which will repeatedly adapt to changes in the milieu or operations and bring the best business outcome. It can also be rapidly, quickly and easily scaled for quick deployment for the other, similar applications [12]. Building a smart home is often an essence for deploying all the sensors, software, network, and physical assets. The data collected and analysis results are shared to the digital twin and can be monitored by an individual [13]. These digital proxies are expected to be built from the domain knowledge of subject matter experts as well as the real-time data collected from the devices [14]. Digital twin is the skill to craft a virtual depiction of the physical elements and the dynamics of how an IoT device operates and device act in response right through its lifecycle [15].

The IoT was motivating the design of digital twins so businesses could take action on data that crosses both the physical and digital worlds. Being able to see it, before DT is build, it has been a long-time aspiration for the manufacturing industry [16]. The technology called Digital Twins which makes it a reality. It allows users to understand how a product would perform before you build. Today's proliferation of sensors, faster computing power and capturing data has grown exponentially [17]. The current acceleration in the usage of digital twins is mostly possible with the Internet of Things and the minor costs of technologies that boosted both IoT and digital twin. This illustrates how a digital twin route sensor generated data from an instrumented advantage and influences replication to forecast malfunction and make out in capabilities [18]. This makes possible an industry to take appropriate action to straight away to correct troubles and optimize the asset's recital [19]. Digital Twin is also called as a replica of an object; it's more than a blueprint or a schematic, virtual twins, shadows, and device virtualization. The process of digital twin creation is shown in Fig. 2.

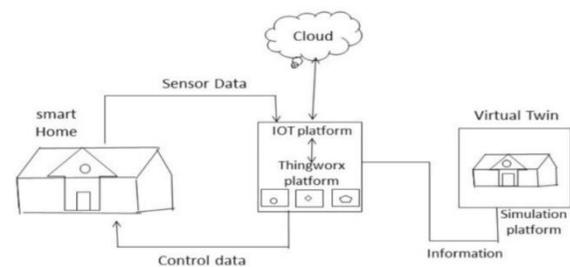


Fig. 2. Digital twin creation.

One of the most important technologies in the Fourth Industrial Revolution is the digital twin, a digital representation of a physical product, service, or production process. Because of the integration of the digital and physical realms, users may now proactively prevent issues through the study of data and the close observation of systems [20]. A digital twin design and system built on a common data standard are described in this research. It describes how to use edge devices, business intelligence, and realistic visualization to create a digital twin for integrated control monitoring [21]. The system's automated generation of digital twin models enables constant communication among field engineers for data collection, designers for modeling, and design engineers for layout changes [22]. As a result of this method, participants can better concentrate on their specific responsibilities in the creation of digital twins.

## II. LITERATURE SURVEY

Due to its critical role in maintaining the availability of several crucial services, the development of safety monitoring and management systems for Critical Infrastructures has attracted increasing attention and concern over the past few years. Due to the unique traits of these systems and the operators' innate reluctance to acts that can cause downtime, this task is difficult. Digital twins can offer a reliable environment for information security or evaluation of prospective mitigation methods to be used in response to certain conditions since they are accurate virtual replicas of physical items or processes. However, one's on-premises implementation can be costly, implying a sizeable CAPEX for whom the return will currently rely on the capacity to intend and deploy an appropriate support infrastructure as well as implement efficient and scalable collection of data and processing mechanisms able to make use of the resources acquired. Sousa et al. [2] proposed an off-premises method for designing and deploying Digital Twins to protect critical infrastructures. In order to help the creation and testing of Machine Learning models to counteract security concerns like Denial-of-Service attacks, such Digital Twins are created utilizing real-time, highly accurate duplicates of Programming Logic Controllers. A significant portion of the predicted CAPEX for the on-premises model was converted like on, pay-as-you-go OPEX through the ELEGANT validation approach, which benefited from the functionality of the Fed4Fire federalized testbeds.

The World Wide Web's structure is comparable to that of the Digital Twin Web, which is made up of conceptual digital twins transmitted through Digital Twin Servers and documented as digital twin findings analyzed by Autiosalo et al. [3]. First before Digital Twin Web can be utilized effectively, standards must be developed; having an easily available server implementation that can encourage the creation of those standards. A Digital Twin Web host built on Git that is open-source and user-friendly is called Twin base. Twin base maintains digital twin papers in a repository, updates them using Git workflows, and then makes them accessible to users through a static web server. Users can view these documents using a library or a standard web browser. The browser interface allows for the free initialization of new server instances. Twin base was created using GitHub, Pages, and Actions, but it can be modified to accommodate different hosting options or self-hosting. To enable the development of derived and alternative server implementations, the author defined Twin base's fundamental architecture. The author offered the idea of a digital twin identity registry in order to answer the need for permanent, openly available, and transferable IDs for the Digital Twin Web. According to performance measurements, depending on the identification registry, the median reaction times for obtaining a digital twin file from Twin base was around 0.4 and 1.2 seconds.

Innovative solutions are required to assure the electrical system's resilience due to the rising frequency of cyberattacks on it. In order to establish a useful framework that can react to various threats on a collection of interconnected micro grids, this work leverages on the development in the Internet of Everything. In order to assure the cyber-physical system's efficient functioning, Saad et al. [6] offered a IoT-based DT of the system that communicates with the control system. The power cyber-physical and DT delivery over the IoT cloud is mathematically formulated. The proposed cybersecurity paradigm, in contrast to others in the literature, can lessen both solitary and group threats. The security protections are put into place utilising cloud computing, and the architecture was evaluated on a distributed system of control. Single-board computers are employed to implement the physical controllers.

In this study, Quan et al. [7] discussed the problem of binary classification for mixed static and dynamic data. The dynamic factors in the novel type of information vary over time and are captured at specific time points, like mixed numerical and categorical data. Due to the significant correlation in each variable caused by this discrete form, more shape and vary over time must be studied, necessitating the urgent need for an effective fusion model. To meet the challenge, the author suggested a novel fusion approach in which the separate findings from variables are converted to function  $f$  via grounds expansion, combined with barely changed via a combination logistic regression model and then the key features are chosen using a group grappling hook penalty term.

The goodness-of-fit criterion has been extensively utilised in the past to assess overall calibration of forecasts. The test aids in determining the significance of inaccurate predictions, which could point to model flaws. As data is created sequentially, the goodness-of-fit test, which is typically

conducted at the conclusion of data collection, may not be able to identify changes in the woman's fit. In order to assess the goodness-of-fit at every time point and give an early warning if major changes take place during model fitting, Qiu et al. [8] looked at the possibilities for employing a new online gawd test.

Sharing data and information from various sources helps scientific partnerships, but maintaining privacy is a top priority. Concern over a possible privacy data leak is growing among researchers, sponsors, and the general public. In order to safeguard the storage and processing for sensitive information in a distributed setting, cutting-edge security techniques have been created. They do not, however, offer any security against information leaking from data analysis results performed by Kim et al. [11]. Studies on differential privacy, a cutting-edge paradigm for privacy protection, have addressed this issue with intriguing findings, but most of them need not be applicable to distributed scenarios. Combining privacy and security procedures is a logical answer to the issue, although simplistic approaches could produce subpar analytical results.

A digital replica of the actual system called as a DT is revolutionizing the way of life. Cyber-Physical Systems (CPS), the IoT, Big Data, Edge-Computing (EC), Artificial-Intelligence (AI), Machine Learning (ML), and other technologies were combined to create DT. DTs are created to improve a variety of applications in business, medicine, smart buildings, smart homes, etc. It's still in the early stages of development. By merging the substantial knowledge on technologies used in the development of DT in industrial and healthcare, this study fills in the gaps analyzed by Khan et al. [12]. The study of DT characteristics, tools and communication technologies used to create DT models, standards and reference models, as well as the researcher's current work in smart factory as well as healthcare, are the main topics of the paper.

A comprehensive range of sensors and actuators, as well as revealing abilities with high-level interface for realistic human contact, enable Pepper, a humanoid robot, to exhibit body language, perceive, and interact with its surrounding environment. In this paper, Cascone et al. [13] described experiences centered on the connection of the digital-twin with both the copies of the smart products in a smart home in order to present the creation of V-Pepper, the Pepper digital replica. Although Pepper robot has hands and arms, its motors and actuator are not strong enough to sustain lengthy testing sessions and training program to learn how to securely handle objects. Here, the metaphor of the digital twin is essential. Machine learning processes can be smoothly transferred to/from the digital-twin with a significant speedup, keeping the physical robot from degrading, by creating a virtual and trustworthy clone of the robot. The given case study provides an inspiration for ambient-assisted functioning in elder care as a practical application. The experience, along with the entire development and design process, has shown that VPepper and the smart cities offer intriguing potential for the physical correctness of the simulation and the accessibility of machine learning tools that may be translated and utilised for actual settings.

### III. PROPOSED METHOD

Home automation includes the use of smart home applications. Instead of just saving power by turning things on and off, smart homes do a lot more [23]. Using these programs gives the impression that users are physically holding and manipulating the virtual things. Users explore using it for things like locking doors, turning off lights and fans, and even controlling the temperature in the fridge [24]. Depending on what users find most useful in the future, users can add on to the functionality of smart homes, transforming the surroundings into a more pleasant and stress-free place to live [25]. The safety and security of useful devices is paramount, and an automated system may help users feel at ease by alerting when to do things allowing users to control who has access to smart devices [26]. Many people find it tedious to constantly check their home gadgets about their working conditions and alerting users about the upcoming issues for hardware devices about their functionality [27].

The advent of digital twin technology can be traced to the development of both virtual technology and data collecting technology [28]. A digital twin is an identical copy of a physical object or person that exists in the real world. The connection and its digital counterpart have multiple possible implementations. Production management, manufacturing, healthcare, smart cities, and other fields all rely heavily on digital technology [29]. Currently, the primary focus of digital twin development is on optimising industrial production. Now that more data can be acquired because to advancements in communication and digitalization technologies, it's time to figure out how to put all that knowledge to good use. As a result, there is a lot of interest in, and momentum behind, the concept of digital twin. All physical entities, including humans, can have their functions monitored, understood, and optimised with the help of digital twins, which also provide constant feedback to enhance quality of life [30]. The ideal way to define a digital twin is as the seamless exchange of information between a real world machine and its digital counterpart. This research introduces the concept of a Digital Twin and explains how it can be applied in various house hold settings as well as the Internet of Things network for better functionality with extended lifetime.

In order to visualize the plans for the twin home, it is necessary to double-check that the planned physical system has successfully received data from sensors, stored it in a database, and run any necessary analytical procedures. An ecosystem was proposed to facilitate the interconnection of IoT devices and sensors, the exposition of the worth of IoT data, the creation of enterprise-level devices, and the authorization of end-users. As the go-between for the sensor and the digital information, and also as the home simulation model, a deep learning model is adopted. The digital duplicate was purpose-built to demonstrate its worth. To begin, a physical assertion is built into smart objects which use sensors to gather information on their current status, work environment, or location. All of the data collected by the sensors is transferred to a central system to be analyzed. This information is analysed in light of existing company metrics and other relevant context details.

Environmental sensors monitor conditions and trigger responses according to those readings in order to cut down on power use and also raise alerts if any prior repair is required. Data collected by sensors revealed the extent to which resources were utilized. The proposed model framework is shown in Fig. 3.

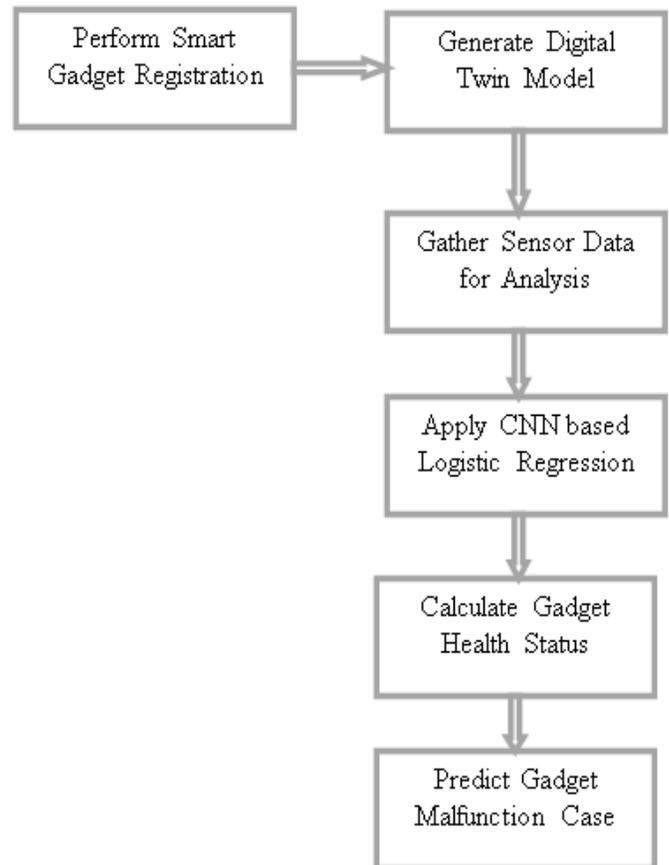


Fig. 3. Proposed model framework.

Digital twin concepts are used to outline a three-stage process for designing a smart home system. Plan, Construct, and Run. The process of making a Digital Copy to begin, a modelling software to produce a digital twin is used. A Digital Twin can be made in a few hours on a computer. Use software to make a virtual model of a Smart Home's infrastructure before constructing the real thing. The software we have here allows us to create a working model of the house. Optimizing device or system performance, reducing unscheduled downtime, and allowing engineers to digitally test solutions before physically repairing something are all possible with digital twins created in this research. The stage of design is critical. Using the IoT platform, the home was equipped with sensors, controllers, and actuators that are linked to a data acquisition component that takes data samples and provides insightful analytical results. This research presents a Deep Convolution Neural Network based Logistic Regression Model with Digital Twins (DCNN-LR-DT) for accurate prediction of smart home gadget functionality levels and to predict the threats in advance.

Algorithm DCNN-LR-DT

**Step1:**

Initially perform the smart gadget registration in the smart home for monitoring the devices performance. The smart gadget registration helps in creating a digital twin with a identity and the process of registration is performed in eq.1.

$$SGD[L] = \sum_{d=1}^N getGDaddr(d) + Node(ID) + TimStamp(getTime()) + Th \quad (1)$$

Here getGDaddr() is used to get the device logical address for further monitoring, Node(ID) represents the physical address to recognize the device and monitoring the malfunctions, Timestamp() is used to get the current time and Th is the threshold value considered. The Smart Gadget Health Status is shown in Fig. 4.

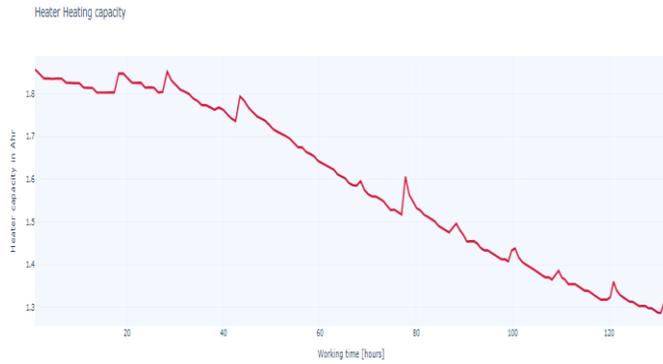


Fig. 4. Smart gadget health status.

**Step2:**

After registration of the gadgets, the digital twin identity is provided and the model generates an accurate digital twin model for gathering the data and analyzing the functionality of the gadget. The digital twin generation is performed in Eq. (2).

$$DigTwin(SGD(L)) = \sum_{d=1}^M getData(SGD(d)) + \sum_{d=1}^M \frac{F(setsimm(SGD(d)))}{count(SGD(d))} \quad (2)$$

Here getData() is used to gather the sensor data and the setsimm() is used to create a simulation model of the gadget using the sensors for monitoring the device functionality. The comparison of physical model is shown in Fig. 5.

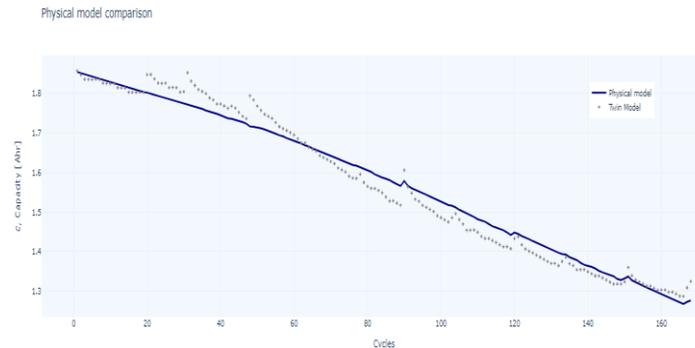


Fig. 5. Comparison of physical model.

**Step3:**

The sensor data is gathered by the central monitoring authority for analyzing the gadgets working process and malfunctions by the smart gadgets and the sensor data analysis is performed in Eq. (3).

$$CS(P) = \frac{\lambda + SGD_{avg} + getsimm(SGD(d, \lambda))}{count(SGD) + \beta} + \sum_{d=1}^M getMax(SGD(d)) \quad (3)$$

$\beta$  is the data interceptions from the sensors, that is the multiple data items collected.  $\lambda$  is the device functionality normal parameter that is set as fixed value. getMax() is used to get the device maximum attribute data for analyzing the device functionality.

**Step4:**

To identify the malfunctions in the device working process, CNN based Logistic Regression model is applied for analyzing a predicting the functionality as in Eq. (4) and Eq. (5).

$$LogReg(f(x)) = \frac{1}{1 + e^{-(x-\mu)}} + \frac{1}{1 + e^{-(\beta_0 - \beta_1 \mu)}} \quad (4)$$

$$MF(SGD(L)) = \sum_{d=1}^L \frac{dStat(LogReg(d))}{\lambda} + \delta(CS(d)) - \tau \quad (5)$$

$\mu$  is the location parameter of the gadget and  $x$  is the scaling attribute,  $\beta$  is the data interceptions from the sensors.  $\lambda$  is the device functionality normal parameter and  $\tau$  is the noisy data gathered from the sensor.  $\delta$  is the model used for highly correlated value set. The error rate between Physical Twin and Digital Twin is shown in Fig. 6.

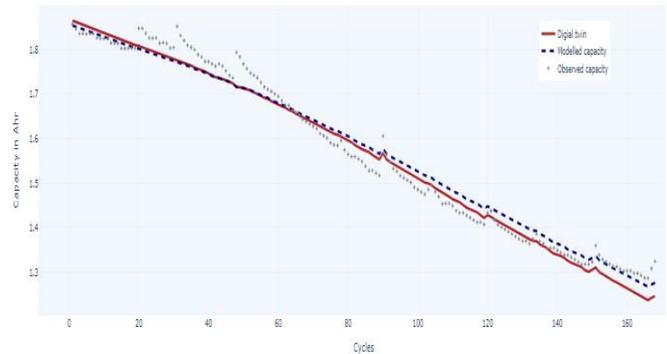


Fig. 6. Error rate between physical twin and digital twin.

**Step5:**

The gadget health status is calculated and updated to the user and the malfunction cause by the sensors is identified and the health status is calculated in Eq. (6).

$$SGDHS(CS(M)) = \sum_{d=1}^M \lambda + \beta * (\maxattr(CS(\lambda)) - \frac{\minattr(\lambda)}{\tau})^2 + \sum_{d=1}^M \lambda * (\tau - \frac{\beta * CS(d)}{count(MF)})^2 \quad (6)$$

Done

#### IV. RESULTS

Digital twin is used to describe physical items that also include digital data. It is also thought of as a technology for symbolizing simulation techniques. The concept of a digital twin is related to other technologies such as cyber physical and digital shadow systems. The connection between these ideas has to be investigated. Designing, running, and repairing the product all make use of digital twin technology. Large volumes of data have been generated by these applications, necessitating a data analysis system for use in fault prediction and maintenance. It is possible to solve the issue with the help of digital twin technology, which acts as a connection between the real and virtual gadgets. Data is the lifeblood of the digital twin concept. Radio Frequency Identification (RFID) tags and readers consist of a variety of elements and sensors. These components are picked and combined so that the digital twin can collect comprehensive data. When it comes to transmitting data to a digital twin via central server, it can be challenging and expensive if the data in question comes in huge volumes and a wide variety.

Data-related technologies, such as data gathering, data mapping, data processing, and data transmission, vary widely depending on the specific use case. To make this data into a digital twin, standard data interfaces are needed. This research presents a Deep Convolution Neural Network based Logistic Regression Model with Digital Twins (DCNN-LR-DT) for accurate prediction of smart home gadget functionality levels and to predict the threats in advance. The proposed model is compared with the traditional Cloud-Based Digital Twinning for Structural Health Monitoring Using Deep Learning (CbDT-SHM) model. The proposed model is implemented in python and executed in Google Colab. The dataset is gathered from the link <https://www.kaggle.com/datasets/prasannaakella/digital-twin-gadget-health>. The results represent that the proposed model performance is efficient than the traditional models.

Equipment like sensors, gadgets, appliances, and other equipment that collect and share data via the web are examples of Internet of Things smart devices. Embedded with other Internet of Things gadgets, they are pre-programmed for certain uses. The smart gadgets in which digital twin will be created will be registered with the model for analysis. The Smart Gadget Registration Time Levels of the proposed and existing models are shown in Fig. 7.

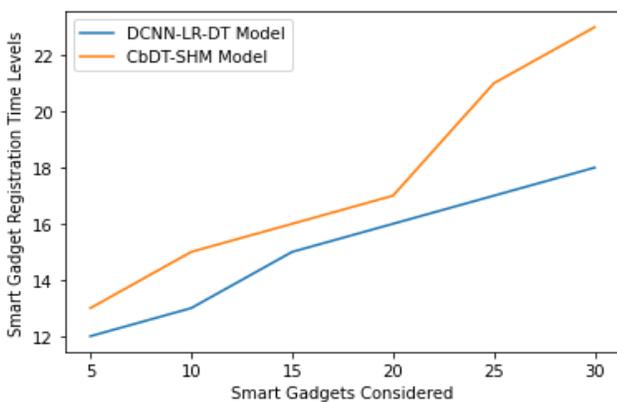


Fig. 7. Smart gadget registration time levels.

For a digital twin to be constructed, information about a physical item or process is needed so that an intangible model can be developed to replicate the actions and states of the physical one. Information collected throughout the product's lifetime may include design documents, production procedures, and engineering blueprints. The Digital Twin Setup Time Levels of the existing and proposed models are compared and the results are shown in Fig. 8.

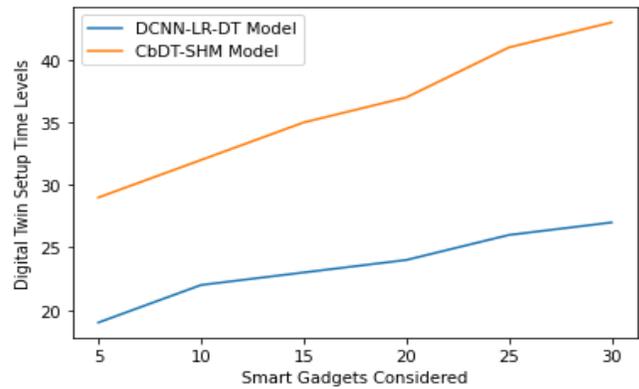


Fig. 8. Digital twin setup time levels.

A digital twin is a digital representation of the physical or system that extends its lifetime, is upgraded from real-time data, and employs simulation, advanced analytics and reasoning to support decision-making. The duplicate can be utilised alongside a prototype to provide input on the product's development or can serve as a model in its own right to mimic what may occur with a tangible version of the product after it is constructed. The accuracy levels of digital twin setup is shown in Fig. 9.

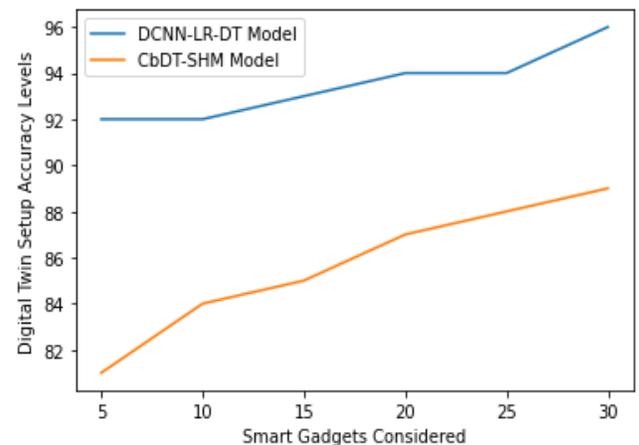


Fig. 9. Digital twin setup accuracy levels.

Data gathered through IoT sensors for processing. Anything from a home thermostat to a motor vehicle could be included in this category. The internet portion of IoT refers to the devices' ability to communicate with one another, share information, and transfer that information across networks for further processing. The sensor data gathered will be used for identification of smart devices working status. The Sensor Data Gathering Time Levels of the proposed and existing models are shown in Fig. 10.

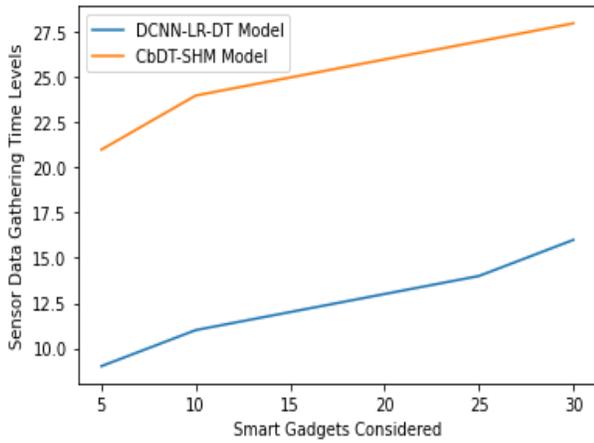


Fig. 10. Sensor data gathering time levels.

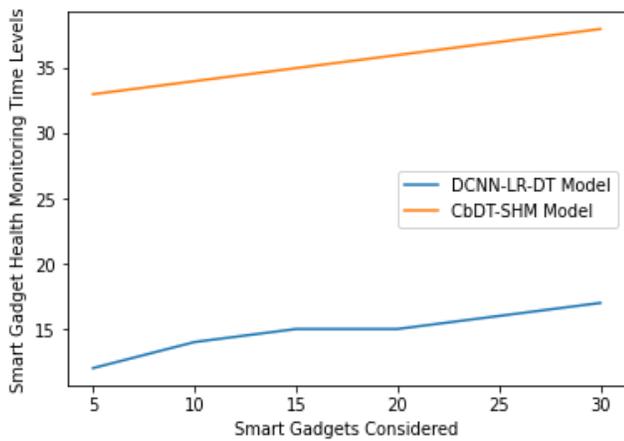


Fig. 11. Smart gadget health monitoring time levels.

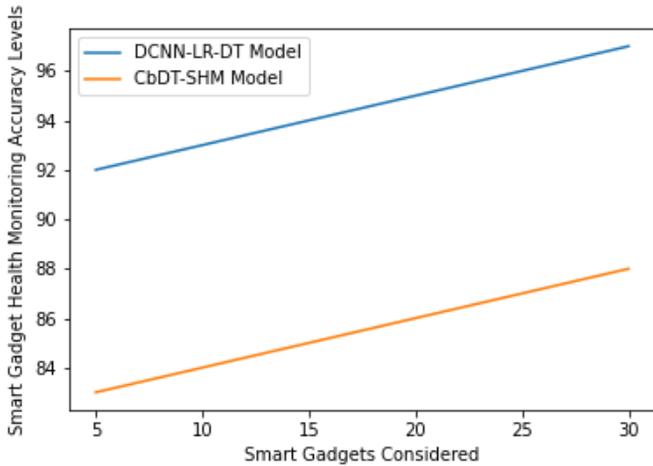


Fig. 12. Smart gadget health monitoring accuracy levels.

In essence, a digital twin is a living representation of a physical talent or system that can continually adjust to new conditions and operations to yield optimal results for the organization. It is also easy to scale up and deploy quickly for use with comparable applications. Deploying all of the sensors, software, networking, and physical assets is crucial to the

construction of a smart building. The concept of a digital twin, or a continuous virtual reproduction of a physical system, has been gaining popularity. It can be used to simulate a smart house, complete with human occupants and robot helpers. Useful in optimizing robotic systems and even the comforts of a smart home for smart gadget health monitoring and suggesting the users for taking necessary actions for long life of gadgets. The Smart Gadget Health Monitoring Time Levels of the proposed and existing models are shown in Fig. 11 and the Smart Gadget Health Monitoring Accuracy Levels comparative results are represented in Fig. 12.

## V. DISCUSSION

The current study explores the potential of digital twins in predicting smart home gadget functionality levels and identifying threats in advance. The proposed model, based on a Deep Convolution Neural Network and Logistic Regression, achieved a high level of accuracy (97%) in identifying hardware risks of smart home gadgets. This study contributes to the growing interest in digital twin technology and its applications in the field of IoT. The results of this study suggest that digital twins can be a valuable tool in improving the efficiency and performance of smart home devices. By providing a virtual representation of physical objects, digital twins can monitor and optimize their functions. Moreover, digital twins can provide constant feedback to enhance the quality of life of individuals using smart home devices. Therefore, the implementation of digital twin technology in smart homes has the potential to significantly enhance the overall user experience.

## VI. CONCLUSION

In the era of IoT and technological advancement, digital twins have emerged as a game-changing invention. The digital twin concept integrates and makes extensive use of cutting-edge technologies such as deep learning, machine intelligence, cloud services systems, big data configurations, software analytics, and the IoT, thereby radically altering IT business efficiency and lowering investment costs. The concept of a digital twin, which bridges the gap between real-world and digital environments, is gaining popularity. The advent of digital twin technology can be traced to the development of both virtual technology and data collecting technology. A digital twin is an identical copy of a physical object or person that exists in the real world. The connection and its digital counterpart have multiple possible implementations. Currently, the primary focus of digital twin development is on optimising industrial production. Now that more data can be acquired because to advancements in communication and digitalization technologies, it is time to figure out how to put all that knowledge to good use. As a result, there is a lot of interest in, and momentum behind, the concept of digital twin. All physical entities, including humans, can have their functions monitored, understood, and optimized with the help of digital twins, which also provide constant feedback to enhance quality of life. The ideal way to define a digital twin is as the seamless exchange of information between a real-world machine and its digital counterpart. This research presents a Deep Convolution Neural Network based Logistic Regression Model with Digital Twins model for accurate prediction of smart home gadget

functionality levels and to predict the threats in advance. The proposed model is limited to check the health status of smart home gadgets. With a focus on various levels, the difficulties and future opportunities of Intelligence DTs in advanced robotics and smart manufacturing can be examined.

#### REFERENCES

- [1] H. V. Dang, M. Tatipamula and H. X. Nguyen, "Cloud-Based Digital Twinning for Structural Health Monitoring Using Deep Learning," in *IEEE Transactions on Industrial Informatics*, vol. 18, no. 6, pp. 3820-3830, June 2022, doi: 10.1109/TII.2021.3115119.
- [2] B. Sousa, M. Arieiro, V. Pereira, J. Correia, N. Lourenço and T. Cruz, "ELEGANT: Security of Critical Infrastructures With Digital Twins," in *IEEE Access*, vol. 9, pp. 107574-107588, 2021, doi: 10.1109/ACCESS.2021.3100708.
- [3] J. Autiosalo, J. Siegel and K. Tammi, "Twinbase: Open-Source Server Software for the Digital Twin Web," in *IEEE Access*, vol. 9, pp. 140779-140798, 2021, doi: 10.1109/ACCESS.2021.3119487.
- [4] M. Alqudah, M. Pavlovski, T. Dokic, M. Kezunovic, Y. Hu and Z. Obradovic, "Fault Detection Utilizing Convolution Neural Network on Timeseries Synchronphasor Data From Phasor Measurement Units," in *IEEE Transactions on Power Systems*, vol. 37, no. 5, pp. 3434-3442, Sept. 2022, doi: 10.1109/TPWRS.2021.3135336.
- [5] M. Yaseliani, A. Z. Hamadani, A. I. Maghsoodi and A. Mosavi, "Pneumonia Detection Proposing a Hybrid Deep Convolutional Neural Network Based on Two Parallel Visual Geometry Group Architectures and Machine Learning Classifiers," in *IEEE Access*, vol. 10, pp. 62110-62128, 2022, doi: 10.1109/ACCESS.2022.3182498.
- [6] A. Saad, S. Faddel, T. Youssef and O. A. Mohammed, "On the Implementation of IoT-Based Digital Twin for Networked Microgrids Resiliency Against Cyber Attacks," in *IEEE Transactions on Smart Grid*, vol. 11, no. 6, pp. 5138-5150, Nov. 2020, doi: 10.1109/TSG.2020.3000958.
- [7] M. Quan, "An Advanced Hybrid Logistic Regression Model for Static and Dynamic Mixed Data Classification," in *IEEE Access*, vol. 10, pp. 73623-73634, 2022, doi: 10.1109/ACCESS.2022.3187767.
- [8] Y. Qiu, L. Liu, X. Lai and Y. Qiu, "An Online Test for Goodness-of-Fit in Logistic Regression Model," in *IEEE Access*, vol. 7, pp. 107179-107187, 2019, doi: 10.1109/ACCESS.2019.2927035.
- [9] D. Zhang, E. Stewart, J. Ye, M. Entezami and C. Roberts, "Roller Bearing Degradation Assessment Based on a Deep MLP Convolution Neural Network Considering Outlier Regions," in *IEEE Transactions on Instrumentation and Measurement*, vol. 69, no. 6, pp. 2996-3004, June 2020, doi: 10.1109/TIM.2019.2929669.
- [10] Z. Zhang and Y. Han, "Detection of Ovarian Tumors in Obstetric Ultrasound Imaging Using Logistic Regression Classifier with an Advanced Machine Learning Approach," in *IEEE Access*, vol. 8, pp. 44999-45008, 2020, doi: 10.1109/ACCESS.2020.2977962.
- [11] M. Kim, J. Lee, L. Ohno-Machado and X. Jiang, "Secure and Differentially Private Logistic Regression for Horizontally Distributed Data," in *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 695-710, 2020, doi: 10.1109/TIFS.2019.2925496.
- [12] S. Khan, T. Arslan and T. Ratnarajah, "Digital Twin Perspective of Fourth Industrial and Healthcare Revolution," in *IEEE Access*, vol. 10, pp. 25732-25754, 2022, doi: 10.1109/ACCESS.2022.3156062.
- [13] L. Cascone, M. Nappi, F. Narducci and I. Passero, "DTPAAL: Digital Twinning Pepper and Ambient Assisted Living," in *IEEE Transactions on Industrial Informatics*, vol. 18, no. 2, pp. 1397-1404, Feb. 2022, doi: 10.1109/TII.2021.3090363.
- [14] R. L. Zhao, D. X. Yan, Q. Liu, J. W. Leng, F. Wan, J. X. Chen et al. "Digital twin-driven cyber-physical system for autonomously controlling of micro punching system" *IEEE Access* vol. 7 pp. 9459-9469 2019.
- [15] M. Fahim, V. Sharma, T. -V. Cao, B. Canberk and T. Q. Duong, "Machine Learning-Based Digital Twin for Predictive Modeling in Wind Turbines," in *IEEE Access*, vol. 10, pp. 14184-14194, 2022, doi: 10.1109/ACCESS.2022.3147602.
- [16] Y. Wu, K. Zhang and Y. Zhang "Digital twin networks: A survey" *IEEE Internet Things J.* vol. 8 no. 18 pp. 13789-13804 Sep. 2021.
- [17] L. Lei, G. Shen, L. Zhang and Z. Li "Toward intelligent cooperation of UAV swarms: When machine learning meets digital twin" *IEEE Netw.* vol. 35 no. 1 pp. 386-392 Jan. 2021.
- [18] J. Lopez, J. E. Rubio and C. Alcaraz "Digital twins for intelligent authorization in the 5G-enabled smart grid" *IEEE Wireless Commun.* vol. 28 no. 2 pp. 48-55 Apr. 2021.
- [19] W. Sun, H. Zhang, R. Wang and Y. Zhang "Reducing offloading latency for digital twin edge networks in 6G" *IEEE Trans. Veh. Technol.* vol. 69 no. 10 pp. 12240-12251 Oct. 2020.
- [20] Y. Lu, X. Huang, K. Zhang, S. Maharjan and Y. Zhang "Communication-efficient federated learning for digital twin edge networks in industrial IoT" *IEEE Trans. Ind. Informat.* vol. 17 no. 8 pp. 5709-5718 Aug. 2021.
- [21] P. Franco, J. M. Martínez, Y. -C. Kim and M. A. Ahmed, "IoT Based Approach for Load Monitoring and Activity Recognition in Smart Homes," in *IEEE Access*, vol. 9, pp. 45325-45339, 2021, doi: 10.1109/ACCESS.2021.3067029.
- [22] N. M. Allifah and I. A. Zuolkernan, "Ranking Security of IoT-Based Smart Home Consumer Devices," in *IEEE Access*, vol. 10, pp. 18352-18369, 2022, doi: 10.1109/ACCESS.2022.3148140.
- [23] M. Wazid, A. K. Das, V. Odelu, N. Kumar and W. Susilo, "Secure Remote User Authenticated Key Establishment Protocol for Smart Home Environment," in *IEEE Transactions on Dependable and Secure Computing*, vol. 17, no. 2, pp. 391-406, 1 March-April 2020, doi: 10.1109/TDSC.2017.2764083.
- [24] M. Marcos Amoroso, R. Moraes, G. Medeiros de Araujo and V. Silva Rodrigues, "Wireless Network Technologies for Smart Homes: A Technical and Economic Analysis," in *IEEE Latin America Transactions*, vol. 19, no. 5, pp. 717-725, May 2021, doi: 10.1109/TLA.2021.9448285.
- [25] M. Khan, J. Seo and D. Kim, "Real-Time Scheduling of Operational Time for Smart Home Appliances Based on Reinforcement Learning," in *IEEE Access*, vol. 8, pp. 116520-116534, 2020, doi: 10.1109/ACCESS.2020.3004151.
- [26] S. Chadoulos, I. Koutsopoulos and G. C. Polyzos, "Mobile Apps Meet the Smart Energy Grid: A Survey on Consumer Engagement and Machine Learning Applications," in *IEEE Access*, vol. 8, pp. 219632-219655, 2020, doi: 10.1109/ACCESS.2020.3042758.
- [27] R. Minerva, G. M. Lee and N. Crespi "Digital twin in the IoT context: A survey on technical features scenarios and architectural models" *Proc. IEEE* vol. 108 no. 10 pp. 1785-1824 Oct. 2020.
- [28] O. E. Marai, T. Taleb and J. Song "Roads infrastructure digital twin: A step toward smarter cities realization" *IEEE Netw.* vol. 35 no. 2 pp. 136-143 Mar./Apr. 2021.
- [29] D. Mitchell et al., "Symbiotic System of Systems Design for Safe and Resilient Autonomous Robotics in Offshore Wind Farms," in *IEEE Access*, vol. 9, pp. 141421-141452, 2021, doi: 10.1109/ACCESS.2021.3117727.
- [30] H. Gong, T. Rooney, O. M. Akeyo, B. T. Branecky and D. M. Ionel, "Equivalent Electric and Heat-Pump Water Heater Models for Aggregated Community-Level Demand Response Virtual Power Plant Controls," in *IEEE Access*, vol. 9, pp. 141233-141244, 2021, doi: 10.1109/ACCESS.2021.3119581.

# A New Privacy-Preserving Protocol for Academic Certificates on Hyperledger Fabric

Omar S. Saleh<sup>1</sup>, Osman Ghazali<sup>2\*</sup>, Norbik Bashah Idris<sup>3</sup>

Studies, Planning and Follow-up Directorate, Ministry of Higher Education and Scientific Research, Baghdad, Iraq<sup>1</sup>

School of Computing, Universiti Utara Malaysia, Kedah, Malaysia<sup>1,2</sup>

Kulliyyah of Information and Communication Technology, International Islamic University Malaysia, Kuala Lumpur, Malaysia<sup>3</sup>

**Abstract**—Academic certificates are integral to an individual's education and career prospects, yet conventional paper-based certificates pose challenges with their transport and vulnerability to forgery. In response to this predicament, institutions have taken measures to release e-certificates, though ensuring authenticity remains a pressing concern. Blockchain technology, recognised for its attributes of security, transparency, and decentralisation, presents a resolution to this problem and has garnered attention from various sectors. While blockchain-based academic certificate management systems have been proposed, current systems exhibit some security and privacy limitations. To address these issues, this research proposes a new Decentralised Control Verification Privacy-Centered (DCVPC) protocol based on Hyperledger Fabric blockchain for preserving the privacy of academic certificates. The proposed protocol aims to protect academic certificates' privacy by granting complete authority over all network nodes, creating channels for universities to have their private environment, and limiting access to the ledger. The protocol is highly secure, resistant to attacks, and allows improved interoperability and automation of the certificate verification process. A proof-of-concept was developed to demonstrate the protocol's functionality and performance. The proposed protocol presents a promising solution for enhancing security, transparency, and privacy of academic certificates. It guarantees that the certificate's rightful owner is correctly identified, and the issuer is widely recognised. This research makes a valuable contribution to the area of blockchain-based academic certificate management systems by introducing a new protocol that addresses the present security and privacy limitations.

**Keywords**—Blockchain technology; hyperledger fabric blockchain; privacy preservation; decentralized control verification privacy-centered (DCVPC) protocol; academic certificates

## I. INTRODUCTION

Academic certificates such as diplomas and transcripts are essential documents that certify an individual's successful completion of a course of study and enable them to pursue diverse employment opportunities within their field [2]. Nonetheless, conventional paper-based certificates are challenging to transport and susceptible to fraudulent activities. As a result, employers and job seekers have lost trust in the verification process, which is now costly and time-consuming.

In response, several institutions have introduced electronic certificates. However, determining authenticity continues to be a prevalent issue. The application of blockchain technology presents a possible solution to this issue by utilising digital

certificates that guarantee authenticity and discourage counterfeiting. Blockchain-based systems like Blockcerts and Block.co have been developed by universities such as MIT and the University of Nicosia (UoN), where students are given control of their own digital credentials and can share them with potential employers [1],[2],[3],[4],[5],[6],[7],[8],[9]. The implementation of decentralisation, peer-to-peer networking, and cryptography in these systems ensure security and immutability. Although the present systems solve the problem of authenticity, they do not tackle other challenges such as fake universities and impersonation.

The objective of this research is to propose a novel protocol based on Hyperledger Fabric to address the challenges associated with managing academic certificates and safeguarding the privacy of identities. In comparison to other blockchain technologies, Hyperledger Fabric provides increased access control and flexibility in protecting privacy. To showcase the efficacy of the proposed protocol, a proof-of-concept will be developed as part of the study.

This paper presents an innovative approach to address a crucial issue in the education sector, which is the secure, transparent, and privacy-preserving management of academic certificates. Academic certificates, such as diplomas and transcripts, play a vital role in enabling individuals to access education and career opportunities. However, traditional paper-based certificates are inconvenient to transport and prone to forgery. Although e-certificates have been developed to address this issue, verifying their authenticity is still a significant challenge. Blockchain technology has emerged as a promising solution due to its features of security, data integrity, transparency, and decentralisation for managing academic certificates.

However, current blockchain-based systems have some limitations when it comes to ensuring security and privacy, which this research aims to address. This study aims to address the following research questions and objectives: How can a protocol based on the Hyperledger Fabric blockchain that is decentralised, privacy-centred and ensures the privacy of academic certificates be developed? Can this proposed protocol enhance the security, transparency, and privacy of academic certificates while facilitating automation and interoperability in the certificate verification process?

This paper introduces the proposed protocol and its implementation and evaluates its functionality and performance, contributing to the development of more secure

\*Corresponding Author.

and privacy-preserving systems for managing academic certificates. The DCVPC Protocol presents a novel and privacy-centred approach to academic certificate verification that utilises the capabilities of the Hyperledger Fabric blockchain. This protocol offers an innovative means of preserving the privacy of academic certificates by enabling decentralised control and verification of these crucial documents. The protocol empowers individuals to regulate access to their certificates, allowing them to disclose only the necessary information. Furthermore, the use of blockchain technology guarantees that certificates are tamper-proof and immutable, enhancing the overall security of the verification process.

The DCVPC Protocol represents a significant advancement in the area of academic certificate verification, providing a secure and innovative solution that has the potential to become an industry standard.

## II. BLOCKCHAIN TECHNOLOGY AND ITS BENEFITS FOR ACADEMIC CERTIFICATES ISSUANCE AND VERIFICATION

The concept of blockchain was first introduced in the Bitcoin white paper in 2008. This distributed ledger leverages consensus and cryptographic techniques to provide a secure and transparent record-keeping system [10],[11],[12],[13],[16]. Since each block contains transactions and a unique hash value, it is difficult to alter or tamper with a block without being detected [16]. Before a transaction can be added to a block in a blockchain network, a consensus must be reached among a group of nodes. As shown in Fig. 1, a block consists of a header and a body, where the body contains the transaction data. The header contains several components, including the Merkle root, a Nonce, a timestamp, and the hash of the previous block. The hash of the previous block is passed to a hash function, which returns a hash value. By recording the hash of the previous block in the current block, the blockchain expands when new blocks are added and linked to it, while also providing an efficient way to detect any attempts at tampering with previous blocks. The timestamp is used to timestamp every newly generated block. The block creation and verification processes only need to be executed once. Merkle trees are binary trees where the labels of non-leaf nodes are the concatenation of the hashes of its child nodes, and the labels of leaf nodes are the hashes of individual transactions in the block body. The Merkle root, also known as the root hash of a Merkle tree, is used to verify the transactions in a block. Instead of verifying each individual transaction in a block, it is sufficient to compare their Merkle root [27]. The structure of a blockchain network is illustrated in Fig. 1, which shows that each block header contains information about the previous block, including its Merkle root, Nonce, timestamp, and hash. The Merkle root refers to the hash of the initial node in a Merkle tree. To further explain the structure of a Merkle tree, let us take the third block as an example of a transaction representation, TX.

The education sector can greatly benefit from the use of blockchain technology due to its various advantages, such as increased security, low cost, improved data access controls, increased accountability and transparency, identity authentication, increased trust, effective student record

management, support for learners' career decisions, and enhanced learner interactivity [28]. The use of blockchain technology in the education sector provides a secure and efficient way of managing academic records and transactions. Due to its decentralised nature, blockchain ensures that only intended recipients have access to shared data or transacted funds, reducing concerns about data safety. The ability to control who can view the saved information is one of the main characteristics and benefits of blockchain. Academic documents such as transcripts, degrees, and student and instructor files can be securely stored, and the blockchain ensures the authenticity of digital certificates and the security of users' identities.

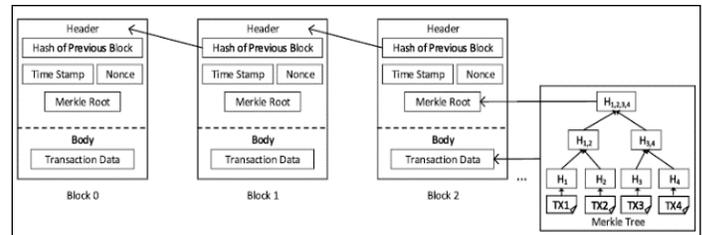


Fig. 1. Blockchain structure.

Blockchain technology also streamlines the process of managing students' personal information, and its adoption in education may reduce the possibility of trading mistakes between parties [28]. Blockchain technology is becoming increasingly important for academic certificate issuance and verification. Traditionally, verifying academic credentials has been a cumbersome and often unreliable process that involved contacting educational institutions and relying on paper records.

The implementation of blockchain technology offers a potential solution for the challenges associated with managing academic certificates. By using a secure and decentralised system, digital credentials can be stored and shared, providing a higher level of trust and transparency in the credential verification process. Utilising blockchain technology allows for tamper-proof certificates to be issued by academic institutions, which can be easily verified by employers and other interested parties. This not only enhances the efficiency of credential verification but also improves the overall integrity of the hiring process.

Furthermore, blockchain technology allows students to exert greater control over their academic records and share them securely and selectively, creating new prospects within the education sector. The implementation of blockchain technology in the issuance and verification of academic certificates has the potential to revolutionise the education industry by enhancing the accuracy and accessibility of academic credentials.

Blockchain technology is widely recognised as a transformative technology for academic certificate issuance and verification. This innovative technology provides a secure, decentralised system for storing and sharing digital certificates, which allows academic institutions to issue tamper-proof credentials that can be easily verified. By leveraging blockchain technology, students can exercise greater control

over their academic records and share them securely and selectively with potential employers and other interested parties.

Moreover, blockchain technology is increasingly recognised as a game-changer for academic certificate issuance and verification. This innovative technology provides a secure, decentralised system for storing and sharing digital certificates, allowing academic institutions to issue tamper-proof credentials that are easily verifiable in real-time, thereby speeding up the hiring process and reducing the risk of fraud. Additionally, the use of blockchain technology enables students to have greater control over their academic records and share them securely and selectively with potential employers or other interested parties. The immutability and tamper-proof nature of blockchain technology ensure that academic certificates cannot be altered or duplicated, providing a higher level of trust and transparency. In summary, the use of blockchain technology for academic certificate issuance and verification offers significant benefits, including greater efficiency, transparency, and security [30],[31].

### III. IMPORTANCE OF HYPERLEDGER FABRIC FOR PRESERVING THE PRIVACY DURING THE PROCESS OF ACADEMIC CERTIFICATES ISSUANCE AND VERIFICATION

The use of Hyperledger Fabric in the process of academic certificate issuance and verification is essential in preserving privacy and security [41]. Academic certificate management involves the exchange of sensitive personal information, making privacy a significant concern. Hyperledger Fabric's architecture enables academic institutions to maintain control over their data, ensuring that private information is not shared with unauthorised parties. The platform provides a secure and private environment where all participants have access to information necessary for the verification process without compromising privacy.

Furthermore, the modular design of Hyperledger Fabric allows for the integration of various identity management systems, providing more precise control over information access. This feature empowers academic institutions to uphold privacy and data security throughout the process of certificate issuance and verification, guaranteeing that sensitive information is only accessible by authorised parties. Consequently, the platform offers a more efficient, reliable, and secure approach to academic certificate management that safeguards the privacy of both students and institutions [25],[26].

Hyperledger Fabric is a highly flexible and scalable platform that enables the deployment of various solutions through a modular subsystem architecture. This feature makes it possible for institutions to scale up to increasingly complex systems. In academic certificate management, Hyperledger Fabric is essential in maintaining privacy and data security during the issuance and verification process. The platform has several key components that work together to preserve privacy. Firstly, its modular architecture allows academic institutions to define their own data access policies and identity management systems, providing control over who has access to sensitive

information. This ensures that only authorised parties can view and verify academic credentials. Additionally, Hyperledger Fabric uses distributed ledger technology to provide a tamper-proof record of all transactions on the network. This feature ensures that certificates cannot be altered or duplicated, which enhances the credibility of the verification process.

Furthermore, Hyperledger Fabric employs a consensus mechanism to ensure that all network participants agree on the authenticity of a transaction before it is recorded on the ledger. This approach guarantees that all participants have a shared view of the network, making it more secure and reliable. Hyperledger Fabric also offers a modular and flexible framework that can be tailored to the unique requirements of various academic institutions. This adaptability enables institutions to integrate their existing systems and processes with the Fabric network, preserving the privacy of their data. Collectively, these components make Hyperledger Fabric an influential tool for academic certificate issuance and verification. It safeguards privacy while ensuring the security and authenticity of academic credentials [29].

The transaction flow in Hyperledger Fabric begins when a client initiates a transaction request by submitting a proposal to the endorsing peer. The endorsing peer then checks the validity of the proposal and simulates the transaction to ensure that it meets the defined rules and regulations. If the transaction is deemed valid, the endorsing peer endorses it by adding a digital signature to the transaction.

Once endorsed, the transaction is sent to the ordering service, which is responsible for ordering the transactions and creating a block. The ordering service ensures that transactions are ordered based on a consensus algorithm and sends the ordered transactions back to the peers. The peers then validate the transaction by checking the endorsement policy and comparing the digital signature of the endorsing peers. If the transaction is valid, the peers commit the transaction to the ledger, making it immutable and tamper-evident.

Hyperledger Fabric also supports private data, which is only visible to parties that have explicit access to it. This is achieved by storing private data off the main ledger and providing access to authorized parties only.

In summary, the transaction flow in Hyperledger Fabric involves several parties, including clients, peers, and orderers, and ensures secure and efficient transactions by utilizing endorsement policies, consensus algorithms, and distributed ledgers. The platform's modular architecture and support for private data make it an ideal solution for enterprises looking to implement blockchain-based systems.

The use of private channels in Hyperledger Fabric enables network participants to have secure and private communication within a subset of the network. The transaction flow in Hyperledger Fabric is carefully designed to ensure the security, scalability, and reliability of the blockchain network, making it an ideal solution for enterprise-grade applications [29]. Fig. 2 illustrates the transaction flow in the Hyperledger Fabric blockchain.

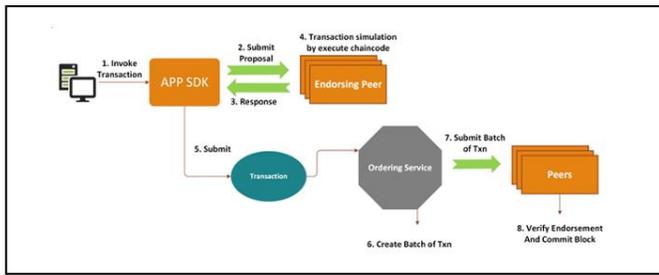


Fig. 2. Transactions flow in hyperledger fabric blockchain.

#### IV. SECURITY AND PRIVACY FEATURES ADOPTED IN HYPERLEDGER FABRIC BLOCKCHAIN

Hyperledger Fabric employs several techniques to ensure the security and privacy of academic certificate issuance and verification processes [18],[33],[34]. Firstly, Fabric employs a permissioned blockchain architecture that restricts access to authorised parties, offering a high level of security against unauthorised access. This feature is particularly crucial in academic certificate management, where the exchange of sensitive personal information necessitates robust security measures. Secondly, Fabric's modular architecture allows for the integration of various identity management systems, enabling academic institutions to maintain control over their data and define their data access policies. This feature is crucial in preserving the privacy of student data, ensuring that private information is not disclosed to unauthorised parties. Thirdly, Fabric leverages distributed ledger technology, which offers a tamper-proof record of all transactions within the network. This feature guarantees the integrity and immutability of academic certificates, ensuring that they cannot be altered or duplicated [24]. This provides a high level of security and authenticity to the certificate issuance process. Fourthly, Fabric's consensus mechanism ensures that all participants in the network agree on the authenticity of a transaction before it is added to the ledger. By achieving consensus, Fabric ensures that all network participants share a unified view of the network, resulting in a more secure and reliable system. Finally, Fabric's private channels enable the creation of secure and private communication within a subset of the network, further preserving the privacy of sensitive data. In summary, the security and privacy preservation techniques utilised in Hyperledger Fabric are well-suited for academic certificate issuance and verification applications. These techniques provide a reliable and secure way for academic institutions to manage certificates while maintaining the confidentiality and data security of all relevant parties [35], [36], [37], [47].

#### V. LITERATURE REVIEW

Blockchain technology offers a potential solution for secure and reliable management of academic certificates due to its distributed ledger architecture and tamper-proof design. Academic institutions can use blockchain technology to issue and verify certificates efficiently and securely. The use of blockchain technology in academic certificate management has gained significant attention in recent years, with various studies exploring its benefits and limitations.

This literature review aims to provide a comprehensive overview of the existing research concerning the use of

blockchain technology for academic certificate issuance and verification. The review will cover various aspects of blockchain technology, including its distributed ledger architecture, consensus mechanisms, privacy features, and security protocols. Additionally, the review will discuss the challenges and opportunities associated with the use of blockchain technology in academic certificate management, and identify areas where future research is needed to further explore this field. The survey aims to offer a thorough comprehension of the present research state concerning the application of blockchain technology in the issuance and verification of academic certificates, and the possible ramifications for academic institutions and stakeholders.

The research [42] explores the potential of blockchain technology in providing a transparent and secure method of recording and maintaining educational certificates and important records. The study highlights the use of digital certificates for evaluating students' academic and extracurricular achievements, and proposes blockchain technology as a secure platform for storing and maintaining them. The research provides an overview of various blockchain-based digital certificate verification systems that employ different authentication techniques and blockchain platforms. It stresses the importance of blockchain technology in ensuring the safety, accessibility, and up-to-date status of digital assets. The study also identifies potential challenges and issues related to academic certification processes in the future. In summary, the research underscores the significance and potential of blockchain technology for academic certificate issuance and verification.

The study [43] introduces a blockchain-based system for the issuance and verification of academic certificates. The system comprises four principal components, namely a verification application with federated identity, an issuing application that involves multi-signature and BTC-address-based revocation, a blockchain, and a local database implemented using MongoDB. The issuing applications manage the primary business logic associated with certificate application, examination, signing, and issuance. They merge the certificate hash with a Merkle tree and send the Merkle root to the blockchain while also handling certificate revocations. The verification application is responsible for verifying the authenticity and integrity of the issued certificates. It includes a web-based page and an Android-based application that retrieves transaction messages through the blockchain API and compares them with the verification data from the receipt. The blockchain acts as a trust infrastructure and a distributed database for storing authentication data, while the MongoDB database manages JSON-based certificates and provides high availability and scalability. Overall, the proposed system leverages blockchain technology to ensure the security and integrity of academic certificates and offers a dependable platform for their issuance and verification.

The study [44] suggests a blockchain-based resolution to tackle the issue of counterfeit educational certificates in Vietnam. The proposed system, referred to as the Vietnamese Educational Certification blockchain (VEcefblock), utilises blockchain technology's features such as anti-forgery information, transaction verification, and smart contracts to

guarantee data transparency and user confidence. The investigation analyses the latest blockchain research and applications to provide insight into the proposed solution. It also presents the development principles of VEcefblock, which involves designing the architecture, business processes, and data mapping structure. Hyperledger Fabric is the blockchain platform used, and the proposed solution is evaluated for performance on the Amazon EC2 cloud. The study underscores the practicality and feasibility of using blockchain technology to address certificate management issues and social problems in Vietnam.

The study [45] introduces a prototype for digital education certificates that uses blockchain technology to enhance the administration and validation of distance education. The prototype is built with a permissioned blockchain, PKI-CA, a digest algorithm, and interactive data authentication via digital signatures. Digital certificates can be issued and verified instantaneously through QR codes or dynamic authorisation codes. Test results demonstrate the prototype's accurate performance with a high throughput of transactions. The proposed system aims to guarantee impartiality and authenticity in education management by ensuring the traceability of student activities and preventing data leakage.

The ongoing research [46] aims to employ blockchain technology to enhance the verification of certificate authenticity. The first stage has led to the development of a prototype, which enables the registration of academic institutions, their faculties, student cohorts, and the issuance of certificate awards. The certificates issued are recorded on the blockchain, ensuring that third parties can verify their authenticity independently of the academic institution, even in the event of its closure. The next stage seeks to expand the prototype to include the registration of medical records, with a focus on ensuring the privacy of sensitive data and granting the owner control over user access to the documents. The final stage involves collecting user and corporate feedback on the proposed prototypes.

MIT Media Lab collaborated with Learning Machine to develop Blockcerts, which uses the Bitcoin blockchain for security. However, this approach slows down transactions, increases prices, and reduces usability. To obtain a diploma using Blockcerts, candidates must install the Blockcerts Wallet software and generate public and private keys. The private keys are stored on users' mobile devices, while MIT receives their public ones. The blockchain stores the diploma's hash value and the date and time it was generated. Graduates can receive digital diplomas that include their public and private keys, which they can use to prove ownership. Graduates can also use the Blockcerts Wallet to share their diplomas with third parties, such as school administrators, future employers, or educational institutions for further education. The system's advantages and drawbacks are discussed in [7], [8], and [11].

There are several benefits for students to use Blockcerts. First, it offers 24/7 access to accredited certification from any location and is valid for the life of the blockchain. Second, students' identities remain private since the blockchain stores only the encrypted hash of their diplomas. Lastly, Blockcerts reduces costs for students by digitising certificates,

empowering schools with greater control over students' academic qualifications, and simplifying the verification process [14].

It has been found that an unauthorised individual could potentially create a fraudulent academic credential using the methods employed by Blockcerts, although [15] found this to be feasible. It is not possible to authenticate the Blockcerts issuing public key. Nevertheless, the use of Blockcerts helps to standardise credentials across universities, streamline the verification process for verifiers, and digitise certificates, all of which have a positive impact on students' time and financial investments.

According to a study [15], one issue with Blockcerts is that it does not provide evidence that the owner of a public key is the authorised issuer. This shortcoming allows unauthorised individuals to produce counterfeit academic credentials that appear to be genuine by impersonating the credential-granting organisation. To tackle this issue, researchers at Birmingham University have proposed a cryptographic digital certificate system called BTCert. BTCert aims to establish a dependable federal ID to verify the legitimacy of the issuing institution, enhance certificate authentication through multiple signatures, and devise a secure revocation method to increase the credibility of certificate revocation.

Birmingham University students and alumni can view their certificates by logging in using their BU credentials. The system allows students to submit their credentials to a third party for verification, and institutional administrators can use it to manage student enrolment, issue certificates with digital signatures, and authorise certificate revocation [6],[11]. Similar to Blockcerts, BTCert generates digital certificates by linking transaction hash values with certificate hashes, and the Merkle root for a certificate set is included in the transaction. The authenticity of the certificate is verified by comparing it to the hash value of the local receipt on the Bitcoin blockchain. The BTCert system comprises a blockchain, an issuing application, a local database, and a verification application. The issuing application is primarily responsible for handling certificate revocation, publishing the Merkle root to the Bitcoin blockchain, and combining the certificate's hash value with a Merkle tree. JSON-based certificates are managed using the local database, while the blockchain stores authentication data [19].

The University of Nicosia was the first institution to use Bitcoin's blockchain to create digital credentials. Students' digital fingerprints are saved on the blockchain, and they can use a verification tool to send a certified PDF version of their diploma to others, avoiding any unnecessary costs. Block.co makes it easy to safely and quickly trade credentials, thus preventing diploma mills. Block.co decides whether to grant a degree after compiling a list of qualified students and assessing a sample application. Diploma forgery is impossible due to the blockchain's digital fingerprinting capabilities, and the degree may be quickly earned and verified by anyone [6],[20].

According to a study [21], a proposed blockchain infrastructure for sharing student information could benefit educators, students, and businesses. The infrastructure is mostly decentralised, as it does not depend on a single server to

access learner data but instead utilises a centralised database. The paper does not provide specifics on implementation or experimental analysis. The authors suggested a blockchain-based infrastructure for archiving student records [22], and the results indicate that storing academic information on the blockchain is more cost-effective than using cloud storage. Smart contracts could utilise access control management to protect users' personal information, which would need to be stored securely with multiple database vendors. However, the study does not provide any concrete outcomes that can be tested or implemented.

According to the authors [23], BcER2 is a database based on blockchain technology for storing academic credentials, including diplomas. The researchers implemented their design using an open-source framework called Hyperledger Composer, as described in the article, which provides a high-level design overview, but no implementation details or test results. Central New Mexico Community College uses blockchain to store student records, enabling students to have greater control over their academic information. Students can download their credentials to any device by using their wallet address. However, their strategy appears to rely on on-chain storage, which is both expensive and inherently unscalable.

#### VI. GAPS AND ISSUES IN TERMS OF PRIVACY-PRESERVING WITH CURRENT BLOCKCHAIN-BASED PLATFORMS IN THE CONTEXT OF ACADEMIC CERTIFICATES ISSUANCE AND VERIFICATION

Although blockchain-based platforms offer several benefits regarding the secure, transparent and decentralised storage and verification of academic certificates, privacy preservation issues remain. One concern is that current blockchain-based platforms may not prioritise privacy and could expose sensitive information to unauthorised parties [32]. Additionally, the public visibility of blockchain transactions and certificates could jeopardise students' privacy and academic records. Furthermore, current platforms may lack reliable mechanisms for identity verification and access control, which can lead to fraudulent activities or unauthorised access to academic records. The non-compliance of current platforms with data protection regulations such as GDPR may also cause legal and ethical problems. Researchers and developers are working on privacy-enhancing solutions, such as the use of zero-knowledge proofs, homomorphic encryption, and multi-party computation, to address these gaps and issues [38], [39], [40], [41]. Therefore, this study aims to suggest a design for the Certificate Verification Control Protocol (DCVPC) based on the Hyperledger Fabric blockchain.

#### VII. RESEARCH METHODOLOGY

The Design Science Research Methodology (DSRM) could be a suitable research methodology for this study. DSRM is widely used in information systems research and concentrates on creating and assessing creative solutions to practical problems [48]. This approach is particularly appropriate for this study since it entails designing and executing a new

protocol based on the Hyperledger Fabric blockchain to protect the confidentiality of academic certificates. The first phase of the methodology defines the problem, which in this case is the lack of privacy in conventional paper-based academic certificates. The second phase involves designing a solution to the problem, which in this case is the proposed DCVPC protocol. The third stage involves creating a model of the suggested solution, which would necessitate constructing and testing the protocol in a real-world setting. Finally, the fourth phase involves assessing the efficiency and effectiveness of the proposed solution. In general, the DSRM methodology could offer a structured approach to developing and evaluating the suggested protocol and could lead to an innovative solution to the issue of preserving the privacy of academic certificates.

#### VIII. THE PROPOSED DESIGN OF CERTIFICATE VERIFICATION CONTROL PROTOCOL (DCVPC)

The initial stage in the design process was to develop a privacy-focused system that allows schools to issue certificates while giving the recipients centralised access to their data. To preserve the confidentiality and integrity of the parties in the network, only approved entities can join the network. However, this does not fully address the importance of decentralisation in this study. Decentralisation ensures that no single entity can exert undue control over the information, and the distribution of power is equalised. Additionally, tracking every alteration can prevent fraudulent and illegal adjustments. Transferring the solution to the Hyperledger network would be the next step towards achieving decentralisation. The previous sections explained why Hyperledger was chosen over competing alternatives. In comparison to other options, Hyperledger's orderer node distinguishes it. Bitcoin and Ethereum, two of the most prominent blockchains, use a consensus form known as probabilistic [25]. In this setting, nodes do not delegate decision-making authority to a central authority; instead, they compete to solve a computational problem  $f(x)$ . When a node successfully solves the problem and adds it to the chain, the probability of the previous block being legitimate increases as the number of blocks in the chain grows, as shown by the expression  $P(\text{block}_{i-1} \text{ valid}) > P(\text{block}_i \text{ valid})$ . This means that the consensus is based on the  $P$  value, which indicates how likely it is that the newly added blocks are valid (Block<sub>i</sub> valid). Hyperledger, on the other hand, uses a deterministic consensus rather than a probabilistic one [28]. When the orderer determines that a transaction should be recorded in the ledger, that decision is conclusive and applies to all affiliated organisations. Each organisation's peers will validate the transaction before sending it to the orderer, who will then package it and send it back to the peers for a final commit.

The proposed protocol comprises several entities, including the ministry, the university, and the student. The ministry is integrated into the system as a hardcoded entity, while universities are represented as organizations, and the channel encompasses the smart contracts. Fig. 3 shows the overall design of the proposed protocol.

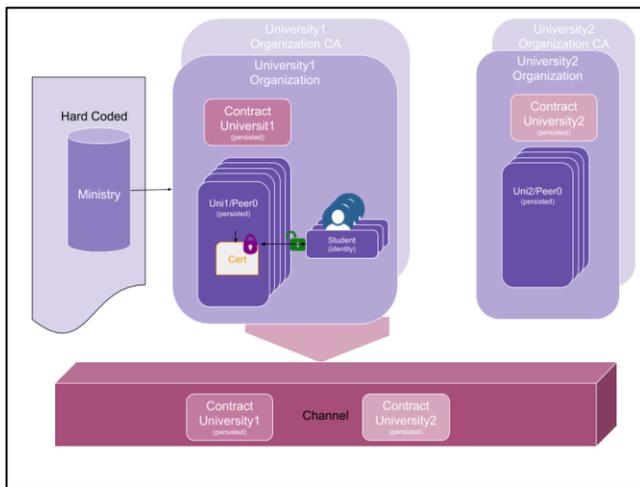


Fig. 3. The desing of the proposed decentralised certification verification privacy control protocol.

Since there is a finite number of ministries, they must be hard-coded into the Hyperledger's inception state. As an alternative, a trustworthy ministry may be hard-coded into the system and would then be able to invite other ministries to join the blockchain, granting them administrative access and allowing them to establish their own affiliated institutions. To maintain confidentiality, a channel is integrated into the suggested architecture.

Knowing one another is a precondition or perhaps a need in educational settings. Anonymity shouldn't play a defining role in the system since it runs counter to the logic of the actual world, where universities should be well-known institutions; it's also crucial to distinguish between anonymity and privacy, which are sometimes confused with one another. It makes no sense that different ministries or universities in other nations have no connection with each other, nor does it make sense that the ministry's own institutions are completely separate. In the actual world, the infrastructure dictates that the entities must be acquainted with one another and work together in some kind.

As a result of the channel implemented by Hyperledger, a relationship can exist among the network's nodes. Assume there are universities under various ministries, and these institutions have partnership programs, such as students taking classes at other universities. In this instance, the suggested protocol can ensure this. Using Hyperledger channels, governments can link together educational institutions so that students, faculty, and researchers from all over the world can easily share data. Hyperledger's architecture makes it possible to propose a solution that preserves privacy on the network and organisational level, making it an excellent place to introduce the protocol. This will ensure that universities worldwide can maintain control over their own data, regardless of which ministry or country is responsible for it. Within the bounds of etiquette, they are free to pursue a romantic partnership in secret. To protect the confidentiality of the network planned for issuing and verifying academic credentials, we present a novel protocol based on the Hyperledger fabric network. The following are the 11 steps of the suggested protocol, and its mathematical algorithm representation is presented in Fig. 4.

Step 1: The ministry, which is the higher authority in the network, creates organisations which are universities.

Step 2: Each ministry creates a channel that connects the universities, authenticated by that ministry.

Step 3: Each university (organisation) creates its own peers under it.

Step 4: Each company's peers maintain their own copy of the ledger and check all transactions before they are permanently recorded.

Step 5: Only universities can host the ledger.

Step 6: The university organization generates Identities for students.

Step 7: The university organization (admin) is the only entity that can issue certificates in the network.

Step 8: All certificates issued are locked in their initial state until the student issues an unlock command.

Step 9: The student can lock and unlock certificates.

Step 10: Third-party entities only need the certificate id, keys and the owner's name to validate and authenticate the certificate.

Step 11: No entity can authenticate a certificate if it is locked.

The suggested protocol gives unrestricted power over the network nodes. By using the channel to merge universities, they can have their own private domain, allowing for information sharing and complete privacy control over each entity. Ministries generate channels to prevent dubious organisations from operating within the network. Universities generate students, which limits the number of random users in the network. In addition, only peers generated by universities are authorised to host the ledger, reducing the risk of an attack and limiting the amount of access. With a decentralised approach, the final commit takes place only when the majority of connected peers approve.

1. MinistryN,  $N \in \{\text{list of countries}\}$ ,  $n(\text{Ministry}) = x$ ,  $x$  is a constant
2. Each MinistryN has  $\{University1, \dots, University \mid i > 0, i = \text{identification of university}\}$
3. MinistryN  $\rightarrow$  ChannelM,  $M = \text{MinistryN}$
4. MinistryN  $\rightarrow$  UniversityMi  $\mid$  UniversityMi  $\in \{\text{MinistryN}\}$ ,  $M = \text{MinistryN}$ ,  $i = \{\text{identification of university}\}$ ,
5.  $\forall$ UniversityMi  $\exists$  MinistryM
6.  $\{\text{UniversityM1}, \dots, \text{UniversityM}\} \subseteq \text{ChannelM}$ ,  $M = \text{MinistryN}$
7. UniversityMi  $\rightarrow$  Peeri,  $\{i \in N\}$
8. Ledger  $\subseteq$  Peer0, ..., Peeri
9. University<sub>admin</sub>  $\rightarrow$  Identities,  $\{s \in \text{Students}\}$
10. University<sub>admin</sub>  $\rightarrow$  Certificates,  $\{s \in \text{Students}\}$
11. Students,  $s \in \text{Students} \rightarrow \text{Lock/Unlock}(\text{Certificates})$

Fig. 4. The mathematical representation of the proposed protocol for preserving the privacy of identities of hyperledger fabric blockchain.

## IX. IMPLEMENTATION AND RESULTS

The implementation of the DCVPC protocol for privacy preservation in a blockchain-based academic certificates management system using Hyperledger Fabric would involve the following steps:

- **Setting up the Hyperledger Fabric network:** This would include installing the necessary software and dependencies, creating the network, and configuring the various components, such as the peer nodes and the ordering service.
- **Developing the smart contract:** The smart contract would be responsible for managing the academic certificates on the blockchain. It would include functions for creating, issuing, and verifying certificates and managing access control.
- **Creating channels:** channels would be created for each university participating in the network. These channels would provide a private environment for each university to manage its own academic certificates.
- **Implementing access control:** Access to the ledger would be restricted and controlled through the use of digital identities and verifiable credentials. Only authorised entities, such as universities and students, could access and make changes to the ledger.
- **Implementing the certificate verification process:** The certificate verification process would be automated using smart contracts. Verifiers would be able to easily access and verify the authenticity of certificates using the public key of the university.
- **Developing the user interface:** A user-friendly interface would be developed for universities and students to interact with the blockchain and manage their academic certificates.
- **Testing and evaluating the proposed protocol:** The proposed protocol would be tested and evaluated using a proof-of-concept to demonstrate its functionality and performance.
- **Deployment:** The final implementation would be deployed on a production network and made available for use by universities and other stakeholders. In the next subsections, the implementation steps are discussed.

### A. Installing the Hyperledger Fabric on the Local System

Setting up a Hyperledger Fabric network involves several steps including the following:

- **Installing the Hyperledger Fabric software:** This includes downloading the Hyperledger Fabric binaries, setting up the necessary environment variables, and installing any additional dependencies such as Go, Docker, and Node.js.

- **Setting up the network:** This includes creating the necessary configuration files for the network, such as the network topology and the configuration of the peer nodes and ordering service.
- **Starting the network:** This includes launching the peer nodes, ordering service, and other components of the network using the command line interface.
- **Joining peers to the network:** After the network is started, other peers can join the network by connecting to one of the existing peer nodes and obtaining the necessary configuration files.
- **Creating channels:** channels can be created by one of the existing peer nodes on the network, and other peers can join these channels by obtaining the necessary configuration files.
- **Installing and instantiating chaincode:** Smart contracts, also known as chaincode, can be installed and instantiated on the network by one of the existing peer nodes.
- **Setting up the SDK:** In order to interact with the network, a software development kit (SDK), such as the Hyperledger Fabric SDK for Node.js needs to be installed and configured.

To set up the Hyperledger Certificate-VPC network, certain prerequisites needed to be fulfilled. The network requires the use of Linux/macOS operating systems to function correctly. The latest version of Hyperledger Fabric (v. 2.2) was installed on the Linux operating system version, as depicted in Fig. 5.

The aforementioned prerequisites were installed successfully by following the official Hyperledger documentation. For the installation, Hyper-V containing Ubuntu 14.04.3 LTS was used on a Windows 10 operating system. Fig. 6 displays a screenshot of Hyperledger Fabric running on the local system.

1. **cURL** — latest version
2. **Docker** — version 17.06.2-ce or greater
3. **Docker Compose** — version 1.14.0 or greater
4. **Golang** — version 1.11.x
5. **Nodejs** — version 8.x (other versions are not in support yet)
6. **NPM** — version 5.x
7. **Python 2.7**
8. **Install Samples, Binaries and Docker Images**

Fig. 5. Prerequisites needed to run the hyperledger fabric.

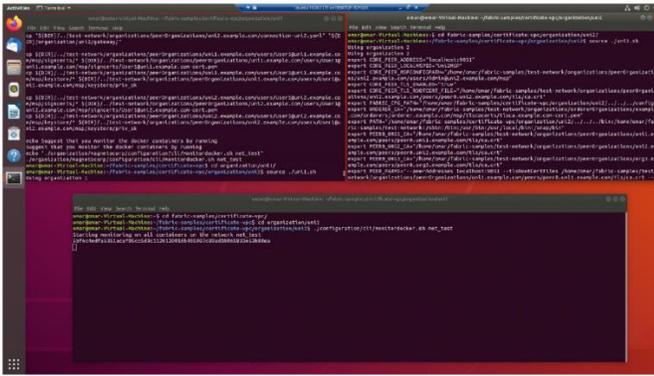


Fig. 6. Screenshot of hyperledger fabric running on the local system.

### B. Hyperledger Certificate-VPC Network

There are several requirements needed to be considered before creating the Certificate-VPC network, and they are as follows:

- Define the network actors.
- Define the peers.
- Define the channel.
- Define the transactions.

The proposed Hyperledger Certificate-VPC infrastructure setup is easily scalable, and the prototype proposed would contain one ministry that creates two universities, and each university creates two users and a peer. The network will suffice for the following:

- Certificate-VPC represents the whole network. One ministry over Hyperledger, which contains two organisations.
- Currently, there are two universities: university1 and university2. There are one or more peers for every organisation.
- A peer can be either a committing peer or an endorsing peer. We set up each organisation with one peer where chaincodes are installed in order to streamline the network configuration. Additionally, this peer commits to verifying transactions within a block.
- Despite the fact that a network may have more than one channel, as shown by the prior design, the Hyperledger Certificate-VPC is only constructed with one channel (the privacy channel).
- A channel is connected to a ledger (blockchain file) to log channel transactions.
- Transactions in Hyperledger Certificate-VPC are issuing, locking, unlocking, and requesting to verify the certificates.

After defining the requirements mentioned previously, setting up the Certificate-VPC network design is described in the next sections.

### C. Setting up the Hyperledger Certificate-VPC Network

To set up the Certificate-VPC network, certain Hyperledger Prerequisites are required, with Linux/MacOS being the preferred operating system. Running the network on Windows can be problematic due to issues with docker. Each main component of the network operates on its own docker. The key components of any Hyperledger network are organisations (in this case, universities) with their respective certificate authority and orderer, which has its own certificate authority. In this study, a ministry with the highest authority in the network will establish two universities/organisations, with the term "university" used to represent an organisation in the Hyperledger to avoid confusion. Uni1 and Uni2 are two hypothetical universities that belong to the ministry in the blockchain's genesis state. The namespace for the Hyperledger network is the URL for the solution, such as CertificateVerificationPrivacyControl.com. To simplify things, example.com is used, and components in the network are reached via a subdomain, such as order.example.com for the orderer and uni1.example.com and uni2.example.com for the universities. Each university has at least one peer represented in its own node as peer0.uni1.example.com. The universities and the orderer have been established, with each university having its own set of users, including students and peers. The administrators possess administrative privileges, whereas the users are granted client access, as demonstrated in Fig. 6. The only missing component is the channel that links these organizations together, as depicted in the illustration provided in Fig. 7. Once the CA server has been set up and all the necessary components have been added to the network, the network is considered to be partially operational. At this stage, two organizations and the orderer have been established, as displayed in Fig. 7. Each organization consists of both users and peers, with administrators possessing administrative privileges, while users are granted client rights.

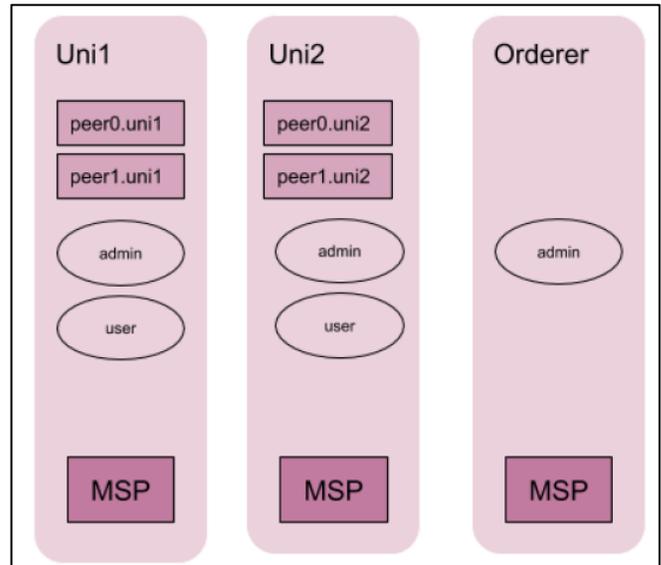


Fig. 7. Components of the created protocol.

With the DCVPC protocol, different types of policies at different layers have to be considered including the following:

1) *Network layer policy*: The policy describes the administrative capabilities of the network, which include the role of the ministry in adding universities and channels as outlined below.

- MinistryN,  $N \in \{\text{list of countries}\}$ ,  $n(\text{ministry}) = x$ ,  $x$  is a constant.
- Each MinistryN has  $\{\text{University}_1, \dots, \text{University}_i \mid i > 0, i = \text{identification of university}\}$ .
- MinistryN  $\rightarrow$  add  $\text{University}_i \mid i > 0, i = \text{identification of university}$ .

2) *Channel layer policy*: This policy outlines the administrative privileges of members at the channel level. This policy permits universities operating under a particular ministry to share a channel and host replicas of the ledger to facilitate their contribution to the network. During the initial phase, all universities created by the ministry will share a single network.

Fig. 8 shows that a channel has been successfully created on Hyperledger Fabric, which includes three organizations represented as universities. The process involved defining the channel configuration, including the policies, orderer settings, and member organizations, where the three universities would be the member organizations. The channel was then created using the Hyperledger Fabric CLI tool or SDK, and during the channel creation process, each university was required to join the channel by creating and signing a certificate and submitting a request to the orderer to join the channel. Once all three universities had joined the channel, they were able to interact with each other and share data, such as academic certificates. This channel provided a secure and private means of communication between the member organizations, ensuring that all transactions were validated and recorded accurately.

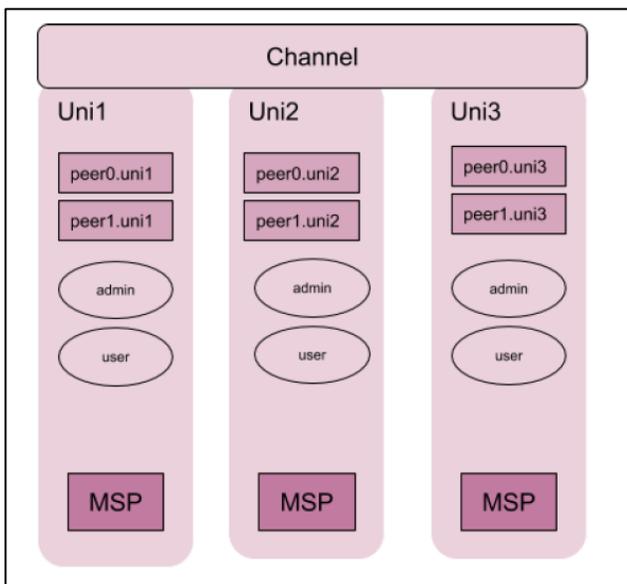


Fig. 8. Channel generation.

3) *Endorsement policy*: The policy being discussed outlines the criteria that must be met to establish the legitimacy of a transaction before it can be recorded on the blockchain. It's important to note that the endorsement policy, if set to something other than the default, would be defined in the configtx.yaml file. By default, the majority of organizations must approve the chaincode before it can be committed to the channel, which will be detailed in the following steps. This policy is sufficient for the design of DCVPC. Once the peers have joined the network, the ledger will consist of three blocks: the initial block created when the channel was established, and two additional blocks for each peer that joins the channel. With this step, most of the proposed design requirements for DCVPC are satisfied.

#### D. Universities Set Up Network starter

To set up the Hyperledger Fabric network, the first step is to generate all required digital certificates, identities of users, and components in the network. This involves using tools such as Hyperledger Fabric's cryptogen or a certificate authority deployed to each organization. To allow entities to use the network, each organization has a Membership Service Provider (MSP) or Certificate Authority (CA) responsible for generating crypto material, which includes private keys and digital certificates. The MSP acts as an identifier for each organization, and all organizations must know each other's MSPs to validate resulting identities. It is essential for university identities to be known and not anonymous, reflecting real-life situations where ministries oversee and monitor universities. Therefore, it is important to identify all file paths for each identity file in the docker-compose configuration file. After generating the crypto material, the necessary components such as organizations, peers, and orderers are started through the use of docker containers. All organizations participating in the network must know each other's MSPs to ensure the validity of transactions [17].

In the design of the blockchain network, the Membership Service Provider (MSP) is essential to identify the organisations participating in the network. An example of this is if an organisation called UUM generates a user named Ahmed using its MSP, other organisations should be able to identify from the signature that Ahmed belongs to UUM, and that UUM is a part of the network. All organisations within the network know each other's MSPs, making it crucial for university identities to be known and not anonymous. This reflects real-life situations where ministries tend to publicise and monitor the universities they oversee. It is also essential for external entities to identify the existing universities. Hiding the identity of universities within the network does not reflect real-life situations, so it was critical to consider this in the design process. The docker-compose configuration file identifies all file paths for each identity file. As shown in Fig. 9, Uni1's identity files are identified in the Docker-Compose configuration file.

```
peer0:uni1.example.com
  container_name: peer0_uni1.example.com
  image: hyperledger/fabric-peer:SIMMAGE_TAG
  environment:
    - MSPID: peer0_mspid
    - CORE_VM_ENDPOINT=unix:///host/var/run/docker.sock
    # the following setting starts chaincode containers on the same
    # bridge network as the peers
    # http://docs.hyperledger.org/2.0/operating/peer-networking
    - CORE_VM_DOCKER_HOSTCONFIG_NETWORKMODE=${COMPOSE_PROJECT_NAME}_test
    - FABRIC_LOGGING_SPEC=INFO
    # Fabric logging spec
    - CORE_PEER_TLS_ENABLED=true
    - CORE_PEER_PROFILE_ENABLED=true
    - CORE_PEER_TLS_CERT_FILE=/etc/hyperledger/fabric/tls/server.crt
    - CORE_PEER_TLS_KEY_FILE=/etc/hyperledger/fabric/tls/server.key
    - CORE_PEER_TLS_ROOTCERT_FILE=/etc/hyperledger/fabric/tls/ca.crt
    # Peer specific variables
    - CORE_PEER_ID=peer0_uni1.example.com
    - CORE_PEER_ADDRESS=peer0_uni1.example.com:7051
    - CORE_PEER_LISTENADDRESS=0.0.0.0:7051
    - CORE_PEER_CHAINCODEADDRESS=peer0_uni1.example.com:7052
    - CORE_PEER_CHAINCODELISTENADDRESS=0.0.0.0:7052
    - CORE_PEER_GOSSIP_BOOTSTRAP=peer0_uni1.example.com:7051
    - CORE_PEER_GOSSIP_EXTERNALENDPOINT=peer0_uni1.example.com:7051
    - CORE_PEER_LOCALMSPID=uni1MSP
  volumes:
    - /var/run:/host/var/run
    - ./organizations/peerOrganizations/uni1.example.com/peers/peer0_uni1.example.com/msp:/etc/hyperledger/fabric/msp
    - peer0_uni1.example.com:/var/hyperledger/production
  working_dir: /opt/gopath/src/github.com/hyperledger/fabric/peer
  command: peer node start
  ports:
    - 7051:7051
  networks:
    - test
```

Fig. 9. Uni1 identity files are identified in the docker-compose configuration file.

In the volumes block, we can observe that the MSP configuration is mapped to the path peers/peer0\_uni1.example.com/msp. When the network is loaded, all the necessary files will be installed in this location. Hyperledger Fabric provides two methods to generate crypto material in the network: using a tool called cryptogen or a certificate authority server. Cryptogen streamlines the identity setup process by automating the generation of crypto material with minimal manual setup. Hyperledger also includes ready-made scripts, such as network.sh, to accelerate the setup process. By executing the network.sh script, the required identities are created using cryptogen and all necessary files are loaded into the corresponding folder path specified in the docker-compose configuration. The command to accomplish this is "cryptogen generate --config=<> --output=<>," with "config" referring to the configuration file defined for each organisation and orderer node inside the cryptogen/\*\_yaml folder. Certain configurations must be in place before starting the network, including each node's configuration in the cryptogen configuration file. The "Count" parameter under "Users" sets the number of users to generate for the university. Once the network.sh script has been executed, the running docker images can be listed, as shown in Fig. 10. Each docker container hosts a specific component in the network.

```
docker images
REPOSITORY          TAG                 IMAGE ID            SIZE
hyperledger/fabric-peer:latest   79617691          79617691            2.8GB
hyperledger/fabric-orderer:latest 79617691          79617691            2.8GB
hyperledger/fabric-peer:latest   79617691          79617691            2.8GB
```

Fig. 10. List of the docker images running.

The Docker configuration file that sets up the network can be found in the file named docker-compose-test-net.yaml. Once the network is up and running, you can access the ledger on either peer using the command docker exec <container-id> peer channel getinfo -c <channel-name>.

By starting the necessary containers and setting up the network, we can observe that the blockchain height is seven blocks, and we will explain the reason behind this shortly. Instead of using the container ID, we can refer to the peer using

its name, which in this prototype includes two organizations, uni1 and uni2, and one orderer. For each organization, we have added specific files, and you can find the list of Hyperledger Fabric files in Fig. 11.

```
ls -ll organizations/peerOrganizations/uni1.example.com/
16:17 ca
16:17 connection-uni1.json
16:17 connection-uni1.yaml
16:17 msp
16:17 peers
16:17 tlscacerts
16:17 users
```

Fig. 11. List of hyperledger fabric files.

The directories for peers and users contain lists of the respective peers and users associated with the organisation. An example of this can be seen in Fig. 12, which displays two main users - Admin and User1.

```
16:17 Admin@uni1.example.com
16:17 User1@uni1.example.com
```

Fig. 12. Users of each organisation.

When we navigate to the users' folders, we can find two sub-folders - msp and tls. The msp folder contains information related to the Membership Service Provider, including the credentials for the user. For instance, for the admin user, we can see the following information inside the msp folder. Additionally, the tls folder contains Transport Layer Security certificates that ensure secure communication between nodes. To get a better understanding of the contents of these folders, refer to Fig. 13, which displays their contents for all users.

```
16:17 admincerts
16:17 cacerts
16:17 config.yaml
16:17 keystore
16:17 signcerts
16:17 tlscacerts
```

Fig. 13. Users of MSP and Tls folder.

The "keystore" directory stores the private key, while the "signcerts" directory contains the certificate for each user. We can determine a user's assigned role by running the command 'openssl x509 -in organizations/peerOrganizations/uni1.example.com/users/User1@uni1.example.com/msp/signcerts/User1@uni1.example.com-cert.pem -noout -subject'. In the case of User1, the role is set as a client. The "ca" directory holds all cryptographic materials, including the private key and certificate for uni1, which can be viewed by running the 'ls'

command. Although "cryptogen" is fast and easy to use, it lacks flexibility when adding or loading entities to the network after deployment. For more control over the certificate generation process, we can use the Fabric CA server. This server hosts the CA, which consists of the private key and CA certificate. To start the network using the Fabric CA server, we can use the same script as before but with the option '-ca'. This will prompt the Fabric-CA-Client tool, which assigns a Fabric CA Admin to manage the addition of entities to the network. To assign the admin for the first university (admin:adminpw are the login details, which can be changed), we can execute the command shown in Fig. 14.

```
fabric-ca-client enroll -u
https://admin:adminpw@localhost:7054 --
caname ca-uni1 --tls.certfiles
${PWD}/organizations/fabric-ca/uni1/tls-
cert.pem
```

Fig. 14. Sample code for fabric-CA-client to assign a fabric CA admin

When adding users to the network, the script './network.sh' calls another script named 'registerEnroll.sh'. This script is responsible for registering and adding users to the network, along with their respective roles. To add a user named Ahmed to the first university, we need to follow these two steps:

Register Ahmed as a client by executing the following code shown in Fig. 15:

```
fabric-ca-client register --caname ca-uni1 --id.name ahmed
--id.secret ahmedpw --id.type client --tls.certfiles
${PWD}/organizations/fabric-ca/uni1/tls-cert.pem
```

Fig. 15. Sample code for registering the users.

Note that if Ahmed were an admin instead of a client, the 'type' parameter would be set to 'admin'.

Generate Ahmed's MSP by executing the code shown in Fig. 16.

```
fabric-ca-client enroll -u
https://ahmed:ahmedpw@localhost:7054 --caname ca-uni1 -M
${PWD}/organizations/peerOrganizations/uni1.example.com/
users/Ahmed@uni1.example.com/msp --tls.certfiles
${PWD}/organizations/fabric-ca/uni1/tls-cert.pem
```

Fig. 16. Sample code for generating the users' MSP.

Once the network is up and running, Ahmed will be successfully registered, as shown in Fig. 17.

```
omar@onar-Virtual-Machine:~/fabric-samples/test-network$ ls organizations/peerOrganizations/uni1.example.com/users/
Admin@uni1.example.com Ahmed@uni1.example.com User@uni1.example.com

Similarly if we add a user Omar to the second university this is what we get:

omar@onar-Virtual-Machine:~/fabric-samples/test-network$ ls organizations/peerOrganizations/uni2.example.com/users/
Admin@uni2.example.com Omar@uni2.example.com User@uni2.example.com
```

Fig. 17. Output of user get registered.

### E. Channel Creation

Creating a channel in Hyperledger Fabric can be easily accomplished by using the same script used to start the network with the command './network.sh createChannel'. In this command, '\$CHANNEL\_NAME' is the variable that sets the channel title. Behind the scenes, a tool called configtxgen is used to create the initial transactions, including the configuration transaction and the peer update transaction for each peer in the intended organizations. These initial transactions prepare the genesis block in the network, which is block #0 and sets the channel. Once the genesis block is set, peers can join the channel using the same block. For example, Fig. 18 displays sample code for creating a channel, while Fig. 19 shows the configuration file of the created channel.

```
peer channel create -o localhost:7050 -c $CHANNEL_NAME
--ordererTLShostnameOverride orderer.example.com -f
./channel-artifacts/${CHANNEL_NAME}.tx --outputBlock
./channel-artifacts/${CHANNEL_NAME}.block --tls --cafile
$ORDERER_CA
```

Fig. 18. Sample code for creating the channel.

```
Configtx.yaml
Profiles:
TwoOrgsOrdererGenesis:
<<< *ChannelDefaults
Orderer:
<<< *OrdererDefaults
Organizations:
- *OrdererOrg
Capabilities:
<<< *OrdererCapabilities
Consortiums:
SampleConsortium:
Organizations:
- *Uni1
- *Uni2
TwoOrgsChannel:
Consortium: SampleConsortium
<<< *ChannelDefaults
Application
<<< *ApplicationDefaults
Organizations:
- *Uni1
- *Uni2
Capabilities:
<<< *ApplicationCapabilities
```

Fig. 19. The configuration file of the created channel.

After the process of setting up the CA server and adding the components to the network, the network is somewhat ready. So to package the smart contract into a chaincode, the command is as follows:

```
peer lifecycle chaincode package unicontr.tar.gz --lang node --  
path ~/contract --label cp_0 ./contract
```

In the above command, the lang specifies the execution language. ~/contract is the path to the smart contract to package. The end result of the above command is a tar.gz file that the admin of that active university can install. The installation is straightforward using the command install. Peer lifecycle chaincode install unicontr.tar.gz . After the installation step is done, the approval process should take place. Each installed chaincode has its own identifier which will allow for determining which chaincode to approve. Because the identifier is a long string, it is easier just to export it as an environment variable. The approval command initiated by the admin is as follows:

```
peer lifecycle chaincode approveformyorg --orderer localhost:7050 --ordererTLSHostnameOverride orderer.example.com --channelID mychannel --name papercontract -v 0 --package-id $PACKAGE_ID --sequence 1 -tl
```

An example of the results of implementing a Certificate Verification Privacy Control Protocol (DCVPC) based on the Hyperledger Fabric blockchain might include:

- Improved security: The use of Hyperledger Fabric blockchain ensures that the certificate issuance and verification process is secure, as all transactions are recorded on a tamper-proof ledger.
- Increased transparency: The CVPC protocol allows for increased transparency during the certificate issuance and verification process, as all transactions are recorded on the blockchain and can be easily audited.
- Reduced fraud: The use of smart contracts and digital signatures in the CVPC protocol can greatly reduce the possibility of fraud, as all certificates are verified and authenticated on the blockchain.
- Improved privacy: The CVPC protocol includes privacy-preserving protocols such as zero-knowledge proofs (ZKP) and homomorphic encryption (HE) to ensure that the personal information of certificate holders is protected during the process of issuance and verification.
- Better interoperability: The use of Hyperledger Fabric blockchain allows for better interoperability among different systems, as the CVPC protocol can communicate with other blockchain networks.
- Automated certificate verification process: smart contracts and blockchain technology automate the certificate verification process and make it more efficient.

## X. DISCUSSION

In this study, we introduce a novel protocol, called Decentralised Control Verification Privacy-Centered (DCVPC), which utilizes Hyperledger Fabric blockchain technology to preserve the privacy of academic certificates. The DCVPC protocol aims to address the limitations of current blockchain-based academic certificate management systems in terms of security and privacy. This is achieved by providing complete authority over all network nodes, establishing private environments for universities, and limiting access to the ledger.

The DCVPC protocol has been designed with a strong emphasis on security, and it is resistant to attacks by restricting access to the ledger and requiring approval from the most connected peers before committing any changes. Additionally, the use of Hyperledger Fabric blockchain technology improves interoperability and automation in the certificate verification process.

We implemented the proposed protocol and developed a proof-of-concept, demonstrating its effectiveness in preserving privacy during the academic certificate issuance and verification process. Our proof-of-concept provided valuable insights into the strengths and weaknesses of the DCVPC protocol and highlighted its potential for preventing forgery and unauthorized access to academic certificates.

One of the significant advantages of the DCVPC protocol is its use of digital identities and verifiable credentials for access control. This ensures that only authorized entities can access and manage academic certificates on the network, which helps to prevent fraud and forgery. Consequently, only deserving individuals can utilize their certificates for education and career opportunities.

In conclusion, the DCVPC protocol has shown promising results in preserving the privacy and security of academic certificates, preventing unauthorized access, and providing a trusted and reliable verification process. By utilizing Hyperledger Fabric blockchain technology and digital identities, our proposed protocol presents a significant step towards achieving a transparent and trustworthy academic certificate management system.

In conclusion, the proposed DCVPC protocol, based on the Hyperledger Fabric blockchain, is a promising solution for improving academic certificates' security, transparency and privacy. Furthermore, we can apply the protocol to other blockchain-based systems to manage educational credentials and enhance it further by incorporating other privacy-enhancing technologies, such as zero-knowledge proofs.

## XI. CONCLUSION

Academic fraud is a significant concern, including both impersonation of certificate recipients and the fabrication of educational institutions. The fake university problem arises when a non-legitimate institution creates a seemingly acceptable academic certificate, while the impersonated receiver problem arises when a person pretends to be the legitimate certificate recipient. Managing authority is also a significant challenge in academic certificate management. Access to resources should be tailored to the responsibilities of

each role. The diploma system involves students, universities, and verifiers, but educational authorities play a crucial role in overseeing institutions at all levels, despite not directly issuing certificates.

This study highlights the importance of Hyperledger Fabric for managing the privacy aspect of academic certificate management systems. We have developed Decentralised Control Verification Privacy-Centered (DCVPC) based on the Hyperledger Fabric blockchain to address these issues. The DCVPC protocol can significantly improve the certificate issuance and verification process by leveraging the security and transparency of the blockchain, as well as privacy-preserving protocols such as zero-knowledge proofs (ZKP) and homomorphic encryption (HE). Additionally, the interoperability and automation of the process provided by the Hyperledger Fabric blockchain can make the process more efficient and streamlined.

It is important to note that the specific results will depend on the particular requirements and constraints of the application, as well as the specific implementation of the DCVPC protocol on the Hyperledger Fabric blockchain. Nonetheless, the DCVPC protocol based on the Hyperledger Fabric blockchain shows great promise in significantly improving the certificate issuance and verification process while preserving the privacy of certificate holders. By addressing the challenges of authority management, academic fraud, and privacy, the DCVPC protocol presents a significant step towards achieving a trustworthy and transparent academic certificate management system.

#### ACKNOWLEDGMENT

This research was supported by Ministry of Higher Education (MoHE) of Malaysia through Fundamental Research Grant Scheme (FRGS/1/2018/ICT04/UUM/02/17).

#### REFERENCES

- [1] KTBS.com (2021). Caddo school employee accused of selling fake diplomas, transcripts. Retrieved July 10, 2021, from [https://www.ktbs.com/news/caddo-school-emplo-ye-accus-ed-of-selli-ng-fake-diplomas-trans-cript/artic\\_le\\_0d8e4eee-e0ea-11eb-ae2a-af6ce6432433.html](https://www.ktbs.com/news/caddo-school-emplo-ye-accus-ed-of-selli-ng-fake-diplomas-trans-cript/artic_le_0d8e4eee-e0ea-11eb-ae2a-af6ce6432433.html).
- [2] Abreu, A. W. S., Coutinho, E. F., & Bezerra, C. I. (2020). A blockchain-based architecture for query and registration of student degree certificates. In Proceedings of the 14th Brazilian Symposium on Software Components, Architectures, and Reuse, 151–160.
- [3] Aini, Q., Rahardja, U., Tangkaw, M. R., Santoso, N. P. L., & Khoirunisa, A. (2020). Embedding a blockchain technology pattern into the QR code for an authentication certificate. *Jurnal Online Informatika*, 5(2), 39–244. Alam, S. (2021). A blockchain-based framework for secure educational credentials. *Turkish Journal of Computer and Mathematics Education*, 12(10), 5157–5167. <https://doi.org/10.17762/turcomat.v12i10.5298>.
- [4] Ataşen, K., & Aslan, B. A. (2020). Blockchain Based Digital Certification Platform: CertiDApp. *Journal of Multidisciplinary Engineering Science and Technology*, 7(7), 12252–12255. From <https://www.jmest.org/wp-content/uploads/JMESTN42353434.Pdf>.
- [5] Caldarelli, G., & Ellul, J. (2021). Trusted academic transcripts on the blockchain: a systematic literature review. *Applied Sciences*, 11(4), 1842.
- [6] Capece, G., Levialdi Ghiron, N., & Pasquale, F. (2020). Blockchain technology: redefining trust for digital certificates. *Sustainability*, 12(21), 8952.
- [7] Castro, R. Q., & Au-Yong-Oliveira, M. (2021). Blockchain and higher education diplomas. *European Journal of Investigation in Health, Psychology and Education*, 11(1), 154–167.
- [8] Chaniago, N., Sukarno, P., & Wardana, A. A. (2021). Electronic document authenticity verification of diploma and transcript using smart contract on Ethereum blockchain. *Register: Jurnal Ilmiah Teknologi Sistem Informasi*, 7(2), 149–163.
- [9] Kshetri, N. (2017). Blockchain's roles in strengthening cybersecurity and protecting privacy. *Telecommunications policy*, 41(10), 1027–1038.
- [10] Vidal, F. R., Gouveia, F., & Soares, C. (2020a). Revocation mechanisms for academic certificates stored on a blockchain. In 2020 15th Iberian Conference on Information Systems and Technologies, 1–6.
- [11] Karamachoski, J., Marina, N., & Taskov, P. (2020). Blockchain-based application for certification management. *Technical Journal*, 14(4), 488–492. <https://doi.org/10.31803/tg-20200811113729>.
- [12] Rahardja, U., Kosasi, S., & Purnama Harahap E., & Aini Q. (2020). Authenticity of a diploma using the blockchain approach. *International Journal of Advanced Trends in Computer Science and Engineering*, 9(1.2), 250-256.
- [13] Bapat, C. (2020). Blockchain for Academic Credentials, from <https://arxiv.org/abs/2006.12665>.
- [14] Baldi, M., Chiaraluce, F., Kodra, M., & Spalazzi, L. (2019). Security analysis of a blockchain-based protocol for the certification of academic credentials, from <https://arxiv.org/abs/1910.04622>.
- [15] Nakamoto, S. (2008). Bitcoin: A peer-to-peer electronic cash system. *Decentralised Business Review*, 21260.
- [16] Wang, Y., & Kogan, A. (2018). Designing confidentiality-preserving blockchain-based transaction processing systems. *International Journal of Accounting Information Systems*, 30, 1–18.
- [17] Hyperledger (2020). A blockchain platform for the enterprise. Retrieved July 10, 2021, from <https://hyperledger-fabric.readthedocs.io/en/latest>.
- [18] Li, R. & Wu, Y. (2018). Blockchain based academic certificate authentication system overview. *IT Innov. Centre, Univ. Birmingham*, 8.
- [19] Block.co (2021). Retrieved July 10, 2021, from <https://block.co>
- [20] Andreev, O., & Daskalov, H. (2018). A framework for managing student data through blockchain. In Proceedings of international scientific conference e-governance and e-communications.
- [21] Han, M., Li, Z., He, J., Wu, D., Xie, Y., & Baba, A. (2018). A novel blockchain-based education records verification solution. In Proceedings of the 19th annual SIG conference on information technology education (pp. 178–183).
- [22] Bessa, E. E., & Martins, J. S. (2019). A blockchain-based educational record repository. *arXiv preprint arXiv:1904.00315*.
- [23] Hope, J. (2019). Give students ownership of credentials with blockchain technology. *The Successful Registrar*, 19(1), 1–7.
- [24] Wang, R., He, J., Liu, C., Li, Q., Tsai, W. T., & Deng, E. (2019). A Privacy-Aware PKI System Based on Permissioned Blockchains. Proceedings of the IEEE International Conference on Software Engineering and Service Sciences, ICSESS, 2018-Novem, 928–931. <https://doi.org/10.1109/ICSESS.2018.8663738>.
- [25] Fabric, H. (2018). A Distributed Operating System for Permissioned Blockchains.
- [26] Brotsis, S., Kolokotronis, N., Limniotis, K., Bendiab, G., & Shiaeles, S. (2020, October). On the security and privacy of hyperledger fabric: Challenges and open issues. In 2020 IEEE World Congress on Services (SERVICES) (pp. 197-204). IEEE.
- [27] Liang, Y. C. (2020). Blockchain for dynamic spectrum management. In *Dynamic Spectrum Management* (pp. 121-146). Springer, Singapore.
- [28] Alammary, A., Alhazmi, S., Almasri, M., & Gillani, S. (2019). Blockchain-based applications in education: A systematic review. *Applied Sciences*, 9(12), 2400.
- [29] Iftekhar, A., Cui, X., Tao, Q., & Zheng, C. (2021). Hyperledger fabric access control system for internet of things layer in blockchain-based applications. *Entropy*, 23(8), 1054.
- [30] K. Verma, R. Singh, and A. Verma, "Blockchain technology for secure and efficient management of academic certificates," *International*

- Journal of Advanced Research in Computer Science, vol. 9, no. 1, pp. 1–7, 2018.
- [31] A. Kshetri, "Blockchain technology for privacy and security in online social networks," *IEEE Communications Magazine*, vol. 56, no. 9, pp. 34–40, 2018.
- [32] Y. Chen, Y. Liu, Y. Zhang, and D. Li, "A privacy-preserving blockchain-based framework for academic certificate verification," *IEEE Access*, vol. 8, pp. 152428–152437, 2020.
- [33] S. H. L. Leong, J. H. M. Lee, and K. W. Chan, "A blockchain-based framework for secure and privacy-preserving academic certificate verification," *IEEE Access*, vol. 8, pp. 57826–57835, 2020.
- [34] J. H. Lee, H. S. Kim, and Y. S. Lim, "A blockchain-based secure and privacy-preserving framework for academic certificate verification," *IEEE Access*, vol. 7, pp. 123361–123369, 2019.
- [35] K. Verma, R. Singh, and A. Verma, "Blockchain technology for secure and efficient management of academic certificates," *International Journal of Advanced Research in Computer Science*, vol. 9, no. 1, pp. 1–7, 2018.
- [36] T. Islam, M. R. Chowdhury, and S. A. R. Hossain, "A blockchain-based secure and privacy-preserving framework for academic certificate verification," *IEEE Access*, vol. 8, pp. 99788–99798, 2020.
- [37] J. Xiong, Y. Li, and X. Shen, "A blockchain-based secure and privacy-preserving framework for academic certificate verification," *IEEE Access*, vol. 8, pp. 141530–141538, 2020.
- [38] Y. Chen, Y. Liu, Y. Zhang, and D. Li, "A privacy-preserving blockchain-based framework for academic certificate verification," *IEEE Access*, vol. 8, pp. 152428–152437, 2020.
- [39] S. H. L. Leong, J. H. M. Lee, and K. W. Chan, "A blockchain-based framework for secure and privacy-preserving academic certificate verification," *IEEE Access*, vol. 8, pp. 57826–57835, 2020.
- [40] J. H. Lee, H. S. Kim, and Y. S. Lim, "A blockchain-based secure and privacy-preserving framework for academic certificate verification," *IEEE Access*, vol. 7, pp. 123361–123369, 2019.
- [41] Saleh, O. S., Ghazali, O., & Rana, M. E. (2020). Blockchain based framework for educational certificates verification. *Journal of critical reviews*, 7(03), 79-84.
- [42] Pathak, S., Gupta, V., Malsa, N., Ghosh, A., & Shaw, R. N. (2022). Blockchain-Based Academic Certificate Verification System—A Review. *Advanced Computing and Intelligent Technologies: Proceedings of ICACIT 2022*, 527-539.
- [43] Awaji, B., Solaiman, E., & Albshri, A. (2020, July). Blockchain-based applications in higher education: A systematic mapping study. In *Proceedings of the 5th international conference on information and education innovations* (pp. 96-104).
- [44] Nguyen, B. M., Dao, T. C., & Do, B. L. (2020). Towards a blockchain-based certificate authentication system in Vietnam. *PeerJ Computer Science*, 6, e266.
- [45] Cheng, H., Lu, J., Xiang, Z., & Song, B. (2020). A permissioned blockchain-based platform for education certificate verification. In *Blockchain and Trustworthy Systems: Second International Conference, BlockSys 2020, Dali, China, August 6–7, 2020, Revised Selected Papers 2* (pp. 456-471). Springer Singapore.
- [46] Curmi, A., & Inguanez, F. (2019). Blockchain based certificate verification platform. In *Business Information Systems Workshops: BIS 2018 International Workshops, Berlin, Germany, July 18–20, 2018, Revised Papers 21* (pp. 211-216). Springer International Publishing.
- [47] Din, I. U., Guizani, M., Kim, B. S., Hassan, S., & Khan, M. K. (2018). Trust management techniques for the Internet of Things: A survey. *IEEE Access*, 7, 29763-29787.
- [48] Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A design science research methodology for information systems research. *Journal of management information systems*, 24(3), 45-77.

# Breast Cancer Prediction using Machine Learning Models

Orlando Iparraquirre-Villanueva<sup>1</sup>, Andrés Epifanía-Huerta<sup>2</sup>,  
Carmen Torres-Ceclén<sup>3</sup>, John Ruiz-Alvarado<sup>4</sup>, Michael Cabanillas-Carbonell<sup>5</sup>  
Facultad de Ingeniería y Negocios, Universidad Norbert Wiener, Lima, Perú<sup>1</sup>  
Facultad de Ingeniería de Sistemas, Universidad Nacional de San Martín, Perú<sup>2</sup>  
Facultad de Ingeniería, Universidad Católica los Ángeles de Chimbote, Perú<sup>3</sup>  
Facultad de Ingeniería, Universidad Tecnológica del Perú, Lima, Perú<sup>4</sup>  
Facultad de Ingeniería, Universidad Privada del Norte, Lima, Perú<sup>5</sup>

**Abstract**—Breast cancer is a type of cancer that develops in the cells of the breast. Treatment for breast cancer usually involves X-ray, chemotherapy, or a combination of both treatments. Detecting cancer at an early stage can save a person's life. Artificial intelligence (AI) plays a very important role in this area. Therefore, predicting breast cancer remains a very challenging issue for clinicians and researchers. This work aims to predict the probability of breast cancer in patients. Using machine learning (ML) models such as Multilayer Perceptron (MLP), K-Nearest Neighbor (KNN), AdaBoost (AB), Bagging, Gradient Boosting (GB), and Random Forest (RF). The breast cancer diagnostic medical dataset from the Wisconsin repository has been used. The dataset includes 569 observations and 32 features. Following the data analysis methodology, data cleaning, exploratory analysis, training, testing, and validation were performed. The performance of the models was evaluated with the parameters: classification accuracy, specificity, sensitivity, F1 count, and precision. The training and results indicate that the six trained models can provide optimal classification and prediction results. The RF, GB, and AB models achieved 100% accuracy, outperforming the other models. Therefore, the suggested models for breast cancer identification, classification, and prediction are RF, GB, and AB. Likewise, the Bagging, KNN, and MLP models achieved a performance of 99.56%, 95.82%, and 96.92%, respectively. Similarly, the last three models achieved an optimal yield close to 100%. Finally, the results show a clear advantage of the RF, GB, and AB models, as they achieve more accurate results in breast cancer prediction.

**Keywords**—Prediction; models; machine learning, cells; breast cancer

## I. INTRODUCTION

Breast cancer can be classified as a type of cancer that occurs in the cells of the breast. Both men and women can get it, although women are more likely than men to suffer from it. The process of breast cancer begins with the uncontrolled

growth of cells in the lining of the breast [1]. At first, there are no symptoms of pain or cancerous growth, and has a low potential for metastatic growth and is limited to the lobe where it grows without generating any symptoms [2],[3]. Symptoms of breast cancer can include anything from a small lump in the breast to changes in the shape of the breast or changes in the color of the skin [4], to identify breast cancer early, it is important to undergo early detection tests, as there are many types of breast cancer and many of them do not cause symptoms at first. Lobular carcinoma in situ, for example, is a type of cancer that occurs in the area of abnormal milk-producing cells of the breast. Invasive lobular carcinoma, which develops in the lobules of the milk-producing mammary glands, people with this symptom experience thickening of the breast tissue, swelling of the breast, and change in skin texture. Ductal carcinoma in situ, this type of cancer usually does not cause symptoms, it is discovered through mammography and invasive ductal is the most common type of cancer accounting for approximately 80% of cases [5]–[7]. There is solid evidence that alcohol consumption, growing older, having dense breasts, family history, radiotherapy treatments, obesity and exposure to radiation increase the risk of breast cancer [2], [8] in turn, it has been shown that prolonged breastfeeding, the development of the physical activity, avoiding harmful consumption of alcoholic beverages and refraining from smoking save, avoiding prolonged use of hormones reduce the risk of breast cancer [8], [9], [10]. Also, mortality from breast cancer in 2020 was 684,996 worldwide, representing 24% of all cancers. While it is true, in recent years the rates of breast cancer events and mortality have been decreasing worldwide [11]. For example, China has the highest rate of breast cancer, with 17.1%; Africa reached 2.5%; the United States at 4%, Japan at 7%; Morocco at 12.5%; Hungary at 2.1% [12]. As shown in Fig. 1, the countries with the highest rates of breast cancer are present in all continents; the continent of Asia concentrates the highest number of deaths from breast cancer.

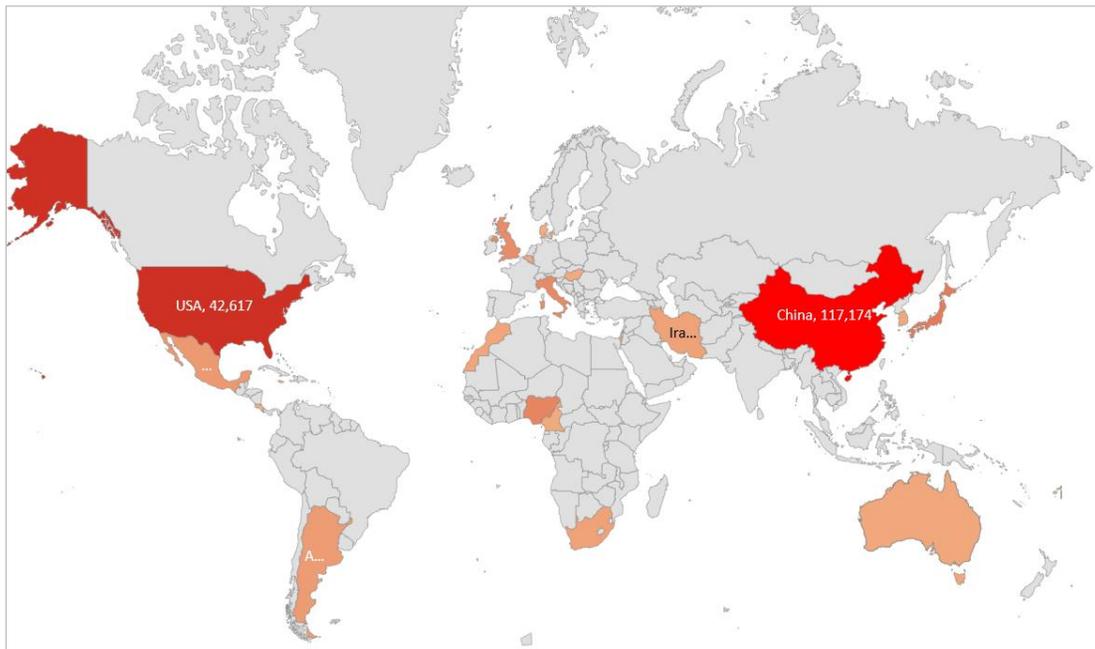


Fig. 1. Breast cancer by country 2022.

In the last decade, technology has undergone impressive development, and with it, ML models are becoming increasingly popular for breast cancer prediction. These models can be used to analyze large patient data sets, such as mammograms, to identify patterns associated with breast cancer development [13]. From these data, ML models can accurately predict a person's risk of developing breast cancer. The accuracy of these models can be further improved by incorporating additional factors such as lifestyle, diet, and family history [14], [15]. With the increasing availability of high-quality datasets and technological advances, ML models are becoming increasingly reliable for breast cancer prediction [16], [17]. There are many types of ML models that can be used to predict the probability that a person will develop breast cancer [18] in this paper we use the classification models such as MLP, KNN, AB, Bagging, GB, and RF, considering that they have excellent performance and performance to analyze and correlate the measurements of the established features. Using features associated with cancer cell imaging, breast cancer can be predicted using ML models. This field of action is in constant development from two deans to after [19], [20].

This paper uses the Wisconsin breast cancer diagnostic dataset to predict and diagnose the likelihood of breast cancer in patients by analyzing six ML models. The dataset is composed of digitized mammogram images and consists of 569 observations and 31 attributes [21]. It also incorporates nine parameters set on a scale of 1 to 10 with values categorized into "benign" or "malignant" tumors.

The article's organization is divided into the following sections. In section II, we describe the most important works that have been done in the area of models of ML. In section III, you will find a description of the method and examples of its application. A summary of the results and discussion of the study can be found in Section IV. Lastly, in Section V, we will present the conclusions that have been reached.

## II. PREVIOUS STUDIES

WHO, American Cancer Society, and scholars have published work related to breast cancer. For example, in [22], [23] they analyzed six ML models with the aim of determining the degree of accuracy of each of them. For this, they used three parameters such as age, cell type with cancer, and cell interface receptors. Also, in [24] developed a predictive model to categorize people with breast cancer using the logistic regression (LR) model, GB model, decision tree (DT), and RF model. Obtaining the following results for the LR model 81.9%; GBT with 82%; RF with 82.8%, respectively. Similarly, in [25] they proposed a model to detect breast cancer using ML models. The tests were performed on a dataset consisting of 317,880 clinical observations. The proposed model achieved an accuracy of 91.22%, and a false rejection rate of 112%. Also, in [26] they used a strategy with feature selection, extraction, and classifier algorithms for breast cancer diagnosis. This study included 762 patients with breast cancer and 138 people without cancer. ML algorithms were used a: 1) LR; 2) SVM; 3) Bagging; 4) GNB; 5) DT; 6) GB; 7) K-NN; 8) BNB; 9) RF; 10) AB, 11) Extra Trees (ET) and 12) MLP. The models that achieved the best results were: LR+MLP with 94%. ML models have demonstrated their contribution to the prediction and early diagnosis of cancer. For example, in [27] they conducted a study to predict and diagnose breast cancer using ML models, for which they used parameters such as specificity, sensitivity, precision, accuracy, precision, and F1 score. The GBDT model obtained a score of 96.77 outperforming all other models. The advancement of Artificial Intelligence (AI) has allowed ML techniques and algorithms to become increasingly efficient in prediction, as evidenced in [28] where they developed a model using ML algorithms to identify and classify different types of cancer. They applied the RF, SVM, and RF models to correctly classify breast cancer cases, obtaining a result: sensitivity of 97.12%, specificity of

96.14%, and accuracy of 97.11%. Artificial intelligence has played a very important role in clinical fields, so much so that, in [29] they evaluated the repeatability of ML model types such as re-regressive, multiclass classification, binary rating, and ordinal classification. The results indicated that classification accuracy improved significantly in most environments. Breast cancer negatively affects the quality of life of patients. In view of this, in [28] they selected an appropriate model to classify and predict the causes that lead to contracting breast cancer, for this purpose they used 970 people with breast cancer. As a result, the SVM model showed the highest sensitivity and an accuracy of 91%, demonstrating that the application of ML algorithms helps the classification of characteristics and the optimization of the genetic algorithm. Accurately distinguishing malignant and benign tumors in patients is crucial to saving lives. That is why in [30] they developed a technique for binary classification of malignant tumors of breast cancer, for which they used three pre-trained convolutional neural network (CNN) models such as RestNet-50, EfficientNetb0, and Inception-v3, applying transfer learning and fine-tuning. The proposed method achieved an accuracy of 98.92%, a sensitivity of 99.87%, a specificity of 97.97%, and an F1 score of 0.9987. In the same line, [30] developed an algorithm based on artificial neural networks (ANN), with the purpose of predicting breast cancer, achieving the following results: accuracy of 98.74%, and an F1 score of 98.02%. Computer-assisted breast cancer screening improves the chances of early detection and diagnosis. So, in [31], [32] proposed a breast cancer screening technique to assess the probability of recurrence of individuals with cancer. The model was trained with 6447 patients diagnosed with breast cancer, the data features were classified with conventional ML and CNN. The best accuracy yielded 88.8%, accuracy 89%, and an F1 score of 0.5. The rapid growth of ML models such as CNNs has promoted the massive use of these technologies in biomedical image classification. For example, in [33] they developed an ML technique to classify breast cancer from histopathological images. The model has been tested with the publicly available BreakHis dataset and has obtained significant accuracy.

### III. METHODOLOGY

This section presents the theoretical basis of the MLP, KNN, AB, Bagging, GB, and RF models and the development of the work to predict and diagnose breast cancer.

#### A. Multi-layer Perceptron

The MLP is an ANN type. It uses backpropagation to train the network [34]. The MLP is composed of multiple layers, each of which is connected to all the others, forming a directed network [35]. The MLP learns a feature from a set of inputs and combines the various features into a set of outputs [36]. The layers usually have weights and polarization units that are adjusted during training. It should be noted that, with the exception of the input nodes, each node in the network is a neuron using a nonlinear activation function, and its equation is given by the following equation and is represented by the following Eq. (1).

$$h_{1j} = f\left(\sum_{i=1}^n w_{ij}x_i + b_j\right) \quad (1)$$

MLP is widely used in supervised learning, where it can learn to classify and predict data. In equation (1),  $h_{1j}$  is defined as node  $j$  of the hidden layer  $h_1$ ,  $w_{ij}$  represents the input gate of node  $j$  of the hidden layer  $h_1$  and  $b_j$  is the bias. In MLP network training, loss functions play an important role. The feature vectors are modeled by the network using loss functions, which are evaluated based on how well the architecture models them. As shown in Fig. 2, the multilayer perceptron model.

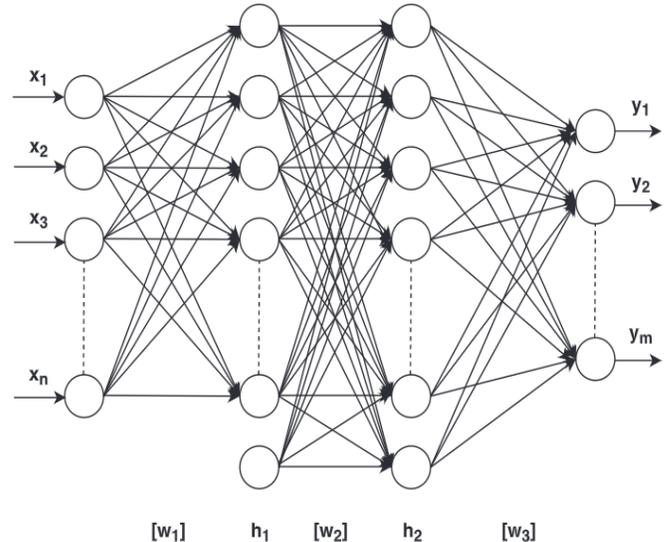


Fig. 2. MLP architecture.

MLPs are limited by their structure, as they are not as flexible as deep learning networks, but they can still be powerful classifiers. Moreover, they do not require large amounts of data, which makes them suitable for many applications [34]. This is the number of training epochs that increases the loss function and gradually reduces its error through optimization.

#### B. K-Nearest Neighbor

As a nonparametric supervised learning classifier, the K-NN algorithm uses proximity to perform classifications and predictions to perform classifications and predictions, respectively [35]. The algorithm stores the attribute vectors and labels used during its training phase so that the algorithm can be retrained [36]. To label the unlabeled vector,  $K$  is set as a user-defined variable, and a label is assigned among the training attributes that are considered most relevant to classify the vector [37]. As for distance metrics for continuous variables, Euclidean distance is used, which is limited to real-valued vectors, for which Eq. (2) is used, and for discrete variables, the overlap metric is used [38].

$$d(x, y) = \sqrt{\sum_{i=1}^n (y_i - x_i)^2} \quad (2)$$

The use of the K-NN model in ML mainly has a better performance in classification and prediction. For example, in data processing, estimating values, automatic recommendations, finance, credit data, in health, its best results have been in predicting the risk of heart attacks, breast cancer, and prostate cancer [39].

### C. AdaBoost

AB is an ML classification algorithm; its principle is based on building strong classifiers by combining basic or weak classifiers. This classification algorithm works on adaptive sampling to select the between samples [40]. This algorithm iteratively trains the weak classifiers, for which it uses weighted data to incorporate it into an ensemble, to then have the strong classifier [41], as shown in Fig. 3.

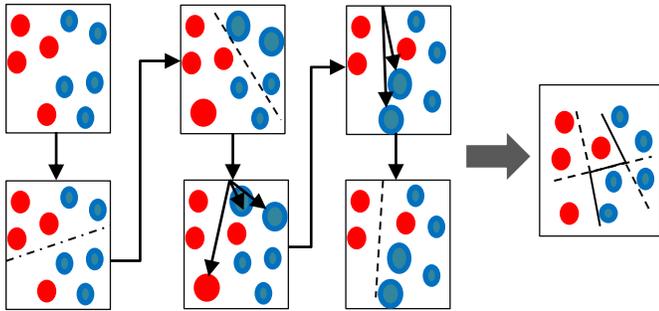


Fig. 3. AB algorithm scheme of work.

Fig. 3 shows that the AB algorithm generates several weak classifiers, where each of the classifiers has a set weight in its performance. Finally, the prediction is obtained by combining the weak classifiers and voting by weight.

### D. Bagging

The bagging model is an ML technique used to improve the accuracy and stability of classification algorithms. It works by combining multiple weak classifiers to form a more robust prediction model [42]. The idea is to create multiple versions of the classifier, each with a different set of parameters, and then combine the results from all of them to produce a better overall prediction [43]. These types of algorithms are run in parallel and seek to take advantage of the independence that exists between single-classifier algorithms, given that the best classifier is chosen by the majority. The Bagging implementation process follows the following steps: Step 1: multiple subsets are created from the data set; Step 2: the base model is created in each of the training subsets; Step 3: each model learns in parallel with each training set; Step 4: the final predictions are determined by combining the predictions of all models.

### E. Gradient Boosting

This classifier combines several weak predictors into a single strong predictor [44]. Using this method, the accuracy of the predictors can be increased by adding predictors sequentially to a set of predictors, each of which corrects the previous one [44]. Basically, the goal of this technique is to find the best predictor for a given problem by iteratively training the model using weak predictors, and gradually improving them until they become strong learners just before solving the problem [45]. This technique has many applications, from data mining to ML or IA.

### F. Random Forest

In the ML field, RF is an algorithm that works as an ensemble. To make predictions, a large number of decision trees are used together to create the decision tree [46]. A decision tree is created using a random subset of the data, and then the results of each tree are combined to make a final prediction, based on the results of all the trees [47]. In terms of classification and regression areas, RF is an extremely powerful algorithm. It can handle large data sets and can be used for both supervised and unsupervised learning. Fig. 4 shows what the model prediction looks like for a new observation.

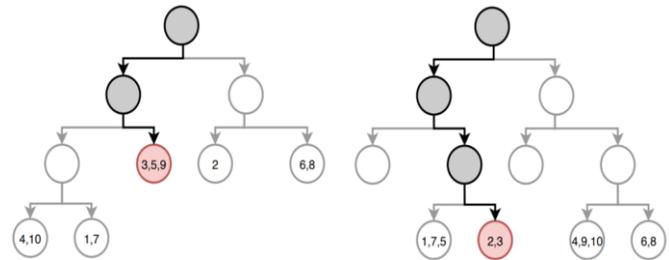


Fig. 4. RF algorithm flowchart.

### G. Understanding Data

According to the American Cancer Society (ACS), a lump is one of the most common symptoms of breast cancer [1]. Several benign breast conditions can cause cancer-like symptoms. However, some of these disappear with time and others require medical treatment. These conditions include cysts, mastitis, hyperplasia, sclerosing adenosis, intraductal papilloma's, fibroadenoma, radial scar, fatty necrosis, and phyllodes tumors. Fig. 5 shows some signs of breast cancer, such as: retracted or inverted nipple, breast or nipple pain, lumps, redness/rash, and changes in skin texture.

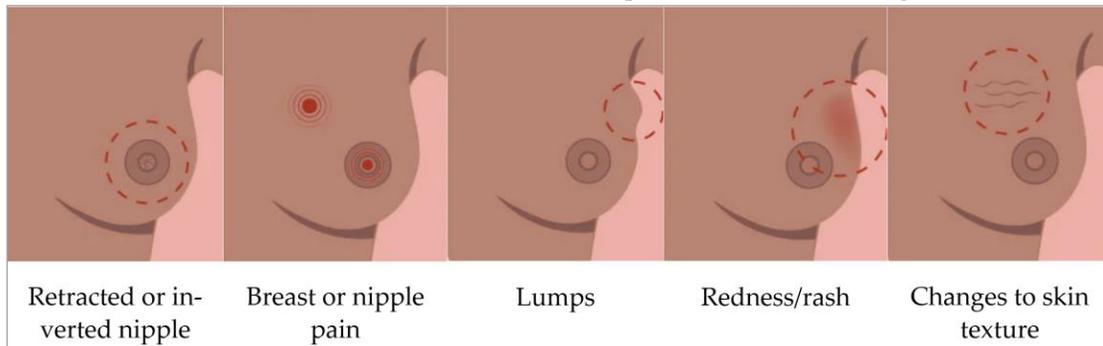


Fig. 5. Signs of breast cancer.

For this work, the Wisconsin Breast Cancer Diagnostic Dataset was used to identify and predict breast cancer. For this purpose, six classification models were used: MLP, K-NN, AB, Bagging, GB, and RF. In addition, univariate analysis, bivariate analysis, and correlation analysis are used for exploratory data analysis (EDA). To evaluate the accuracy of the model, the following methods are used: confusion matrix, classification report, and AUC. The dataset corresponds to digitized images of samples and is composed of 569 observations and 31 attributes: diagnosis, Radius-mean(R-ME), Texture-mean(T-ME), pe-perimeter-mean (P-ME), area-mean(A-ME), smoothness-mean(S-ME), compact-ness-mean(C-ME), concavity-mean(CO-ME), concave points-mean(CP-ME), sym-metry-mean(S-ME), fractal dimension-mean(FD-ME), radius-se(R-SE), tex-ture-se(T-SE), perimeter-se(P-SE) area-se(A-SE), smoothness-se(S-SE), compact-ness-se(C-SE), concavity-se(CO-SE), concave points-se(CP-SE), symmetry-se(S-SE), fractal-dimension-se(F-D-SE), radius-worst(R-WO), texture-worst(T-WO), perimeter-ter-worst(P-WO), area-worst(A-WO), smoothness-worst(S-WO), compact-ness-worst(CO-WO), concavity-worst(C-WO), concave points-worst(CP-WO), sym-metry-worst(S-WO) and fractal-dimension-worst(F-D-WO).

H. Data Cleansing

The data cleaning process, for this case study, was performed using Python programming language was performed using a variety of libraries and techniques. Among the libraries used were Pandas, NumPy, SciPy, Scikit-learn, and NLTK. The Pandas library was used to read data, clean it and manipulate it. It is useful for dealing with missing values, outliers, and other problems. The NumPy library was used to perform calculations on the data, such as mean, median, mode and standard deviation. SciPy and Scikit-learn are declared for the use of ML and statistical analysis. Also, it is used to perform regression, clustering, and other types of analysis. NLTK library is declared for further use for data processing. Also, it will be used to extract text features, such as sentiment analysis and keyword extraction. We then proceeded with loading the dataset and identifying each of the variables, as shown in Table I. The number of variables and the type of data for each of the variables. In addition, in this section, we try to eliminate all duplicate data, handle outliers and deal with incorrect data.

TABLE I. DATASET VARIABLES AND DATA TYPES

Column	not empty Count	Dtype
[diagnosis]	569 (not empty)	Blob
[R-ME]	569 (not empty)	Float 64
[T-ME]	569 (not empty)	Float 64
[P-ME]	569 (not empty)	Float 64
[A-ME]	569 (not empty)	Float 64
[S-ME]	569 (not empty)	Float 64
[C-ME]	569 (not empty)	Float 64
[CO-ME]	569 (not empty)	Float 64
[C-P-ME]	569 (not empty)	Float 64
[S-ME]	569 (not empty)	Float 64
[FD-ME]	569 (not empty)	Float 64

[R-SE]	569 (not empty)	Float 64
[T-SE]	569 (not empty)	Float 64
[P-SE]	569 (not empty)	Float 64
[A-SE]	569 (not empty)	Float 64
[S-SE]	569 (not empty)	Float 64
[C-SE]	569 (not empty)	Float 64
[CO-SE]	569 (not empty)	Float 64
[CP-SE]	569 (not empty)	Float 64
[S-SE]	569 (not empty)	Float 64
[F-D-SE]	569 (not empty)	Float 64
[R-WO]	569 (not empty)	Float 64
[T-WO]	569 (not empty)	Float 64
[P-WO]	569 (not empty)	Float 64
[A-WO]	569 (not empty)	Float 64
[S-WO]	569 (not empty)	Float 64
[CO-WO]	569 (not empty)	Float 64
[C-WO]	569 (not empty)	Float 64
[CP-WO]	569 (not empty)	Float 64
[S-WO]	569 (not empty)	Float 64
[FD-WO]	569 (not empty)	Float 64

I. Exploratory Data Analysis

EDA is an approach to data analysis for organizing key features. Primarily, EDA is used to see what the data can say beyond the formal task of modeling or hypothesis testing. EDA is also used to check the data for interesting features or outliers that may suggest the need for further examination. In addition, EDA can be used to evaluate the assumptions of a model before fitting it to the data. In order to visualize the data graphically, the diagnosis column first had to be enumerated so that Malignant(M)=1, Benign(B)=0. Then, the ID column was set for the dataset, the ID column will not be used for ML. For this, the countplot(), plt.figure() and print() functions were used. As shown in Fig. 6.

Now, for a better understanding of the content of Table II, it is important to have basic knowledge about variance, standard deviation, number of samples, or the maximum and minimum values. This type of information provides a better understanding of what is happening with the data. Therefore, before visualization understands standardization, feature extraction, and feature selection.

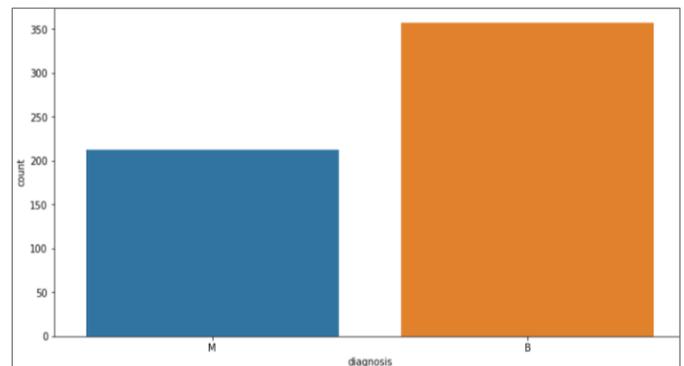


Fig. 6. M and B cancer diagnosis count.

TABLE II. STANDARDIZATION, EXTRACTION, AND SELECTION OF CHARACTERISTICS

	Radius mean	medium texture	Perimeter mean	middle zone	Smoothness mean	Compactness mean
count	[569.00000]	[569.00000]	[569.00000]	[569.00000]	[569.00000]	[569.00000]
mean	[14.127292]	[19.289649]	[91.969033]	[654.889104]	[0.0963600]	[0.1043410]
std	[3.5240490]	[4.3010360]	[24.298981]	[351.914129]	[0.0140640]	[0.0528130]
min	[6.9810000]	[9.7100000]	[43.790000]	[143.500000]	[0.0526300]	[0.0193800]
25%	[11.700000]	[16.170000]	[75.170000]	[420.300000]	[0.0863700]	[0.0649200]
50%	[13.370000]	[18.840000]	[86.240000]	[551.100000]	[0.0958700]	[0.0926300]
75%	[15.780000]	[21.800000]	[104.10000]	[782.700000]	[0.1053000]	[0.1304000]
max	[28.110000]	[39.280000]	[188.50000]	[2501.00000]	[0.1634000]	[0.3454000]

For better visualization of the data, we used the seaborn library, but we classified the features into three groups because the differences between the feature values were so high that it was impossible to observe them, as shown in Fig. 7. Each group includes 10 features for a more effective presentation of the data.

Fig. 7 can be seen. For example, that the T-ME features, the median of M and B appear separate, so it can be very useful for classification. The FD-ME feature, however, does not separate the median of the M and B, so the median in this case cannot be used to classify the data. For reasons of space, the following groups are not shown. In the classification, it was also shown that the variables C-WO and CP-WO are very similar. However, it cannot be stated that they are correlated with each other, in the case of being correlated; one of the two variables is eliminated. To compare the two characteristics more deeply, the joint plot is used.

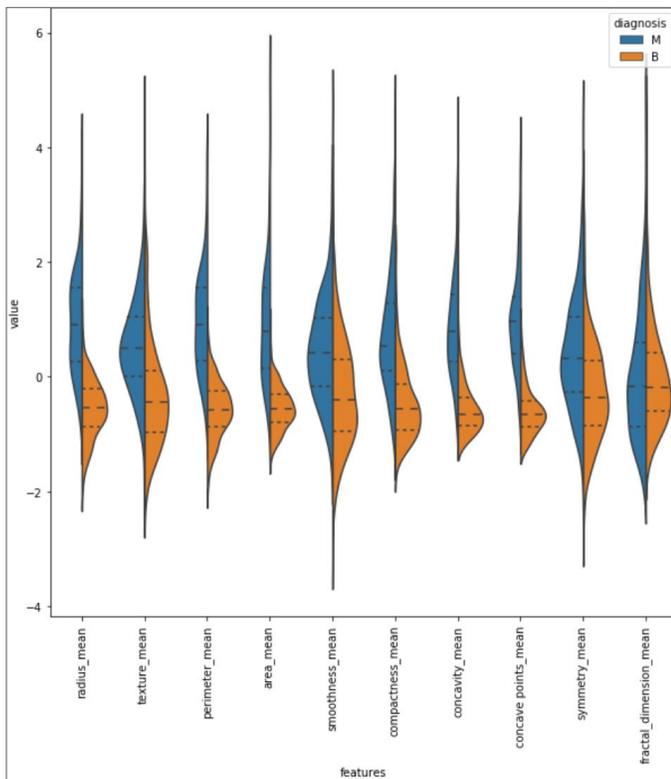


Fig. 7. Standardization and classification of characteristics.

In the next step, features are selected using correlations, univariate features are selected, recursive feature elimination with cross-validation is performed, and attribute categorization is performed. MLP, K-NN, AB, Bagging, GB, and RF classification are used to train the model and predict. As shown in Fig. 8, the R-ME, M-ME, and A-ME features are correlated with each other, so only the A-ME feature will be used. In this way, the features that are correlated are found, with support of the classifiers. C-ME, CO-ME, and CP-ME are correlated with each other, so only CO-ME is chosen. In addition, R-SE, P-SE, and A-SE are correlated, so only A-SE was used. R-WO, P-WO, and A-WO are correlated, so I use A-WO. CO-WO, C-WO, and CP-WO are correlated, so C-WO was used. C-SE, CO-SE, and CP-SE are correlated, so I use CO-SE, T-ME, and T-WO are correlated so I use T-ME, A-WO, and A-ME are correlated so I use A-ME. Specifically, X and Y are not correlated at all; the correlation seen in Fig. 8 is such a strong correlation by chance.

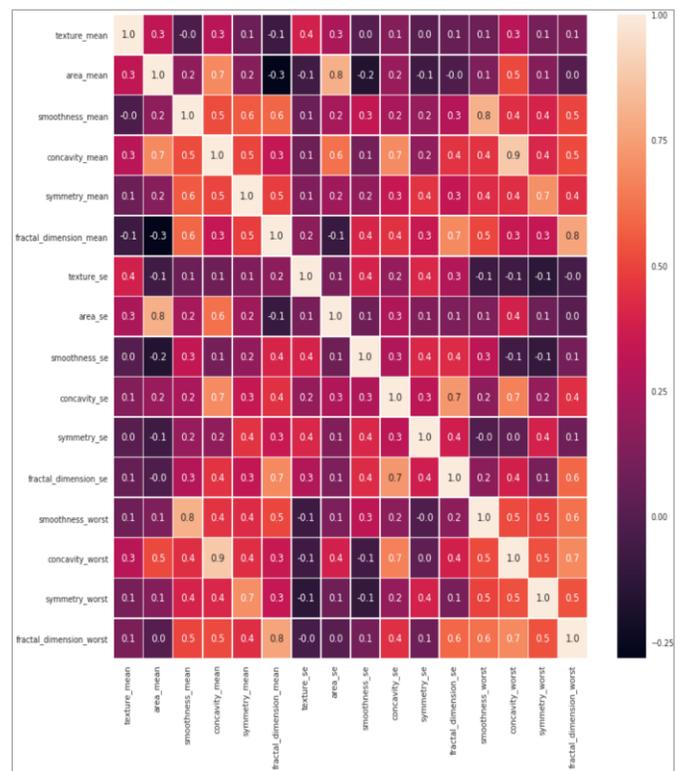


Fig. 8. Numerical correlation of variables.

As part of this work, we use the normalization technique for feature scaling to convert feature values into a mean-centered distribution with unit standard deviation, and this feature scaling method has been widely used in ML algorithms. There are several types of neural networks, such as MLP, K-NN, AB, Bagging, GB, and RF. For example, there is a requirement to normalize features in algorithms such as K-NN and MLP. As a result of the different properties measured by the dataset at each scale, there are heterogeneous features among the datasets at each scale.

**J. Model Training and Testing**

In univariate feature selection, SelectKBest is used which eliminates all features except those with high scores. This method allows choosing the number of features to use. For example, the number of features(k)=5, which means that the model must find the 5 best features, this is achieved with the following function: SelectKBest(arg, k=5).fit(x\_train, y\_train). The results are presented in Table III.

The next step consists of preparing the MLP, K-NN, AB, Bagging, GB, and RF models for training and validation using the train\_test\_split(), project\_data.drop(),

X\_train.select\_dtypes() and Pipeline() functions. The latter allows training the model with the data by adjusting its parameters to create a model that can accurately predict the result while evaluating the model to ensure its accuracy and reliability.

Then the prepare\_model() function is used to compile the model with a given number of features. It takes the features as an argument and returns a compiled model as its output. Also, the function prepare\_confusion\_matrix(y\_true, y\_pred) is used to print the confusion matrix, as shown in Fig. 9. Similarly, the function prepare\_classification\_report() is used to generate the classification report for the given results. Finally, the prepare\_roc\_curve() function allows preparing the receiver operating characteristic (ROC) curve and calculates the false positive rate and the true positive rate, which allows for measuring specificity, and sensitivity, among others. After the evaluation, the following results were obtained [Bagging: 99.78021978021978%; K-NN: 96.7032967032967%; RF: 100.0%, AB: 99.56043956043956%; GB: 100.0% and MLP: 96.26373626373373626%]. It should be noted that only four models have been presented in Fig. 9: Bagging, K-NN, AB, GB.

TABLE III. SELECTION OF UNIVARIATE CHARACTERISTICS

list: [				
6.06916433e	3.66899557e	1.00015175e	1.30547650e	1.95982847e
3.42575072e	4.07131026e	6.12741067e	1.32470372e	6.92896719e
1.39557806e	2.65927071e	2.63226314e	2.58858117e	1.00635138e
1.23087347e]				
List of features: Index([				
texture_mean	area_mean	smoothness_mean	concavity_mean	menmetry_fraction_mean
texture_rease	suavidad_se	concavidad_se	simetría_se	fractal_dimension_se
suavidad_peor	concavidad_peor	simetría_peor	fractal_dimension_peor	)]

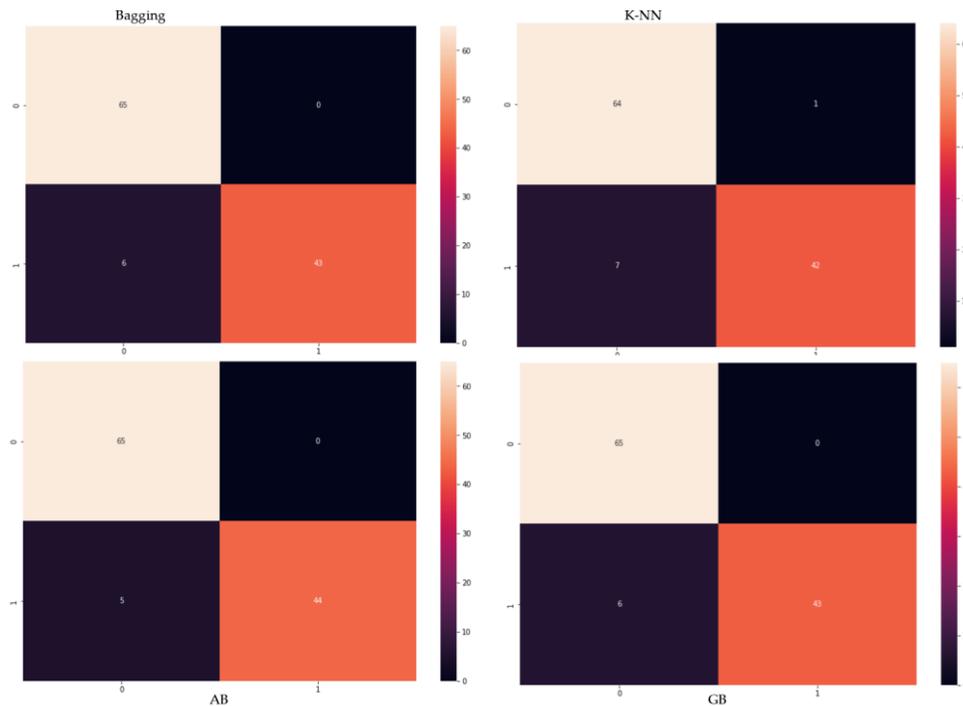


Fig. 9. Matrix of confusion.

#### IV. RESULTS AND DISCUSSION

After training the MLP, K-NN, AB, Bagging, GB, and RF models, on the data set, a learning algorithm is created and used for training. The performance of the models with

unobserved data is then evaluated. The evaluation of each of the models was performed by testing their performance on unseen data. Metrics such as accuracy, precision, recall, F1 score, and ROC curve are used to determine model performance as shown in Table IV.

TABLE IV. MODEL EVALUATION RESULTS

<b>bagging classifier Report</b>				
	<b>accuracy [%]</b>	<b>recall [%]</b>	<b>f1-score [%]</b>	<b>support</b>
B	93	97	95	71
M	95	88	92	43
accuracy			94	114
macro avg	94	93	93	114
weighted avg	94	94	94	114
<b>KNN classifier Report</b>				
	<b>accuracy [%]</b>	<b>recall [%]</b>	<b>f1-score [%]</b>	<b>support</b>
B	92	99	95	71
M	97	86	91	43
accuracy			94	114
macro avg	95	92	93	114
weighted avg	94	94	94	114
<b>RF classifier Report</b>				
	<b>accuracy [%]</b>	<b>recall [%]</b>	<b>f1-score [%]</b>	<b>support</b>
B	93	93	93	71
M	88	88	88	43
accuracy			91	114
macro avg	91	91	91	114
weighted avg	91	91	91	114
<b>AB classifier Report</b>				
	<b>accuracy [%]</b>	<b>recall [%]</b>	<b>f1-score [%]</b>	<b>support</b>
B	97	92	94	71
M	87	95	91	43
accuracy			93	114
macro avg	93	93	93	114
weighted avg	93	93	93	114
<b>GB classifier Report</b>				
	<b>accuracy [%]</b>	<b>recall [%]</b>	<b>f1-score [%]</b>	<b>support</b>
B	96	94	95	71
M	91	93	92	43
accuracy			94	114
macro avg	93	94	93	114
weighted avg	94	94	94	114
<b>MLP Report</b>				
	<b>accuracy [%]</b>	<b>recall [%]</b>	<b>f1-score [%]</b>	<b>support</b>
B	95	97	96	71
M	95	91	93	43
accuracy			95	114
avg	95	94	94	114
weighted avg	95	94	94	114

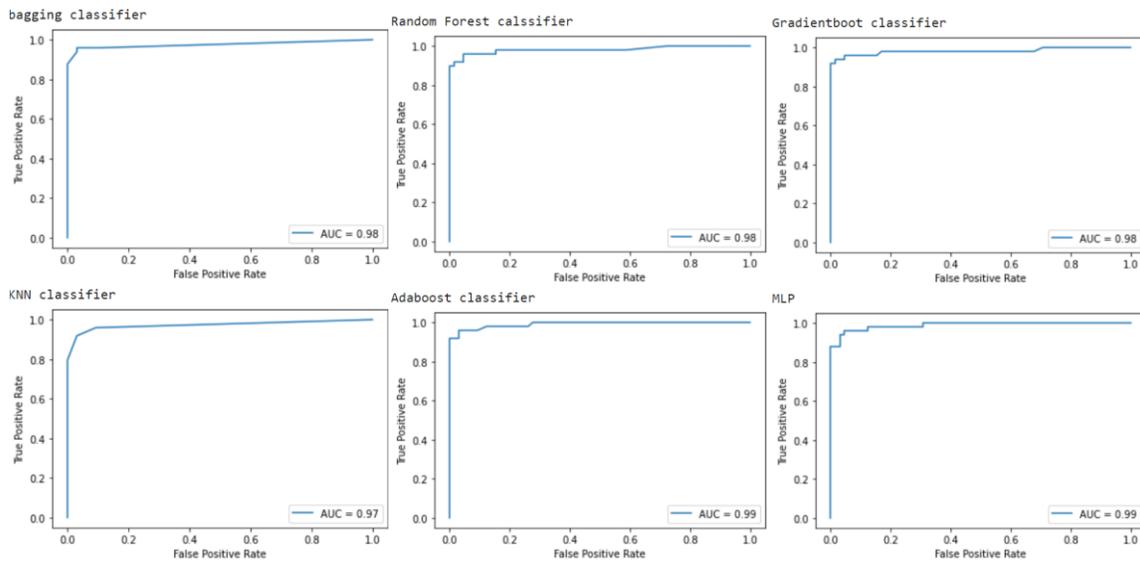


Fig. 10. Performance curve of ML models.

The false positive rate for each model is calculated in a similar way. For example, in the case of the Bagging technique, it helps to improve stability and accuracy by creating various models; in the K-NN model it is the ratio of false positives to the total number of points; in the RF model, it is the rate of false positives that the classifier incorrectly identifies a negative event (e.g., a "no" or a "0") as a positive event (e.g., a "yes" or a "1"); in the AB model the "false positive" rate and the true positive rate depend on the algorithm, the complexity of the data analyzed and the parameters used for the classifier. In general, the false positive rate is quite low, and the "true positive" rate is quite high for AB and in the MLP model, the false positive rate is the probability of misclassifying a true negative case as a positive case. In much the same way the true positive rate is calculated. For example, in the Bagging model, the true positive rate is the proportion of correctly classified positive cases divided by the total number of positive cases; similarly, in the RF and GB model, the true positive rate is the rate at which the classifier correctly identifies a positive event (e.g., a "yes" or a "1"); in the MLP model, the true positive rate is the probability that it correctly classifies a true positive case. Now, for each model, we evaluated the AUC (AUC) performance curve. For example, the models used in this work; Bagging, K-NN, RF, AB, GB, and MLP, obtained the following performance: 98%, 97%, 98%, 98%, 99%, 98%, and 99%, respectively. Fig. 10 shows that the performance curve of each of the models is optimal, reaching practically on average 98%, this makes it possible to opt for any of the models used in this work to classify and predict breast cancer.

For the training and validation of each model used, we worked with an adequate data set. The results shown in Fig. 10 and Table IV show that the performance of each model was successful in cancer prediction accuracy. These results showed superiority in the same ML models in [24] and [26] where the Bagging and K-NN models achieved a performance of 96.47% and 96.40% in predicting Breast Cancer. These results do not determine that one is better than the other, on the contrary, the

performance rate varies according to different factors, and one of them is the volume of data with which it is trained. On the other hand, in [26] they developed a model to predict breast cancer, for which they used the RF model, with which they achieved an accuracy performance of 97.1%, very similar to the 98% accuracy obtained in this work. AI has played a very important role in clinical fields, and models such as AB have contributed a great deal in this field, since it is the model that has achieved the best results, reaching 99% in this study. Likewise, in [29] in the binary classification of malignant tumors of breast cancer, it reached 99.92% accuracy, which makes it the best model for classifying and predicting breast cancer. Similarly, the GB model, which is an excellent classifier by adding predictors sequentially, achieved a 98% performance rate in training, which is also in agreement with the results obtained in [30], where they used the GB model for the purpose of predicting breast cancer, where it achieved a 98.74% performance rate. Finally, the MLP model is characterized as one of the best predictors, this predictor learns a feature from a set of inputs and combines the different features in a set of outputs, the performance rate of this model has been 99%, and it is a result with a high pre-accuracy rate, which allows it to be a reliable option for the prediction of breast cancer. Also, [20], [21] used this model with three clinical factors: age, cancer cell type, and cell surface receptors, obtaining satisfactory results, with a performance rate of 98%. The summary of the analysis of the 6 models used in this work to predict breast cancer is presented in Table V.

TABLE V. SUMMARY OF THE ANALYSIS

Model	Train Accuracy	AUC SCORE
Bagging	99.56	0.97
KNN	95.82	0.97
RF	100	0.98
Adaboost	100	0.96
GB	100	0.97
MLP	96.92	0.98

## V. CONCLUSIONS

Prediction of different types of cancer is one of the most complex fields of medical engineering and AI. In this work, 6 ML models were trained for breast cancer prediction, for which the Wisconsin breast cancer diagnostic dataset was used, with the purpose of predicting and diagnosing in patients the probability of having breast cancer. The dataset corresponds to digitized images of samples and is composed of 569 observations and 31 attributes. Also, the performance of the results of each of the models was analyzed, as shown in Fig. 10. Also, the behavior was compared in the context of the work developed: the Random Forest classifier, Adaboost, and Gradientboot, achieved the best results of 100%, more accurate in terms of breast cancer prediction. The normalization technique was used for feature scaling with the purpose of converting the feature values into an input distribution at the mean with a unit standard deviation. This can be seen in the numerical correlation of variables in Fig. 8, also, Table III shows the univariate characteristics. Table V shows the accuracy of each model: Bagging 99.56%; KNN 95.82%; Random Forest 100%; Adaboost 100%, Gradientboot 100%; and MLP 96.92%. The main contributions of this work consist of the evaluation of 6 ML models for breast cancer prediction. Likewise, the results keep a clear originality of this work, and at the same time confirm that the results obtained in this work are related to other similar works that used ML techniques applied to breast cancer prognosis.

In the future, a possible development that would complement the use of the models would be the development of a mobile application based on services to consume the implemented model. The most important contribution of this work is that doctors through ML models can analyze the data of breast cancer patients in a personalized way to predict their effectiveness, constituting a support tool for health. Limitations of this work include: 1) The data used for training may be biased, which means that there may be biases between terms; 2) the quality of ML model data depends on the quality and volume; if the data is limited, the results will be inaccurate; 3) in terms of resources, training ML models requires a processor with a high responsiveness.

## REFERENCES

- [1] American Cancer Society, "What Is Breast Cancer?," 2020. <https://www.cancer.org/cancer/breast-cancer/about/what-is-breast-cancer.html>.
- [2] World Health Organization, "Breast cancer," 2021. <https://www.who.int/news-room/fact-sheets/detail/breast-cancer> (accessed Jan. 02, 2023).
- [3] World Cancer research Fund International, "Breast cancer statistics International." <https://www.wcrf.org/cancer-trends/breast-cancer-statistics/>.
- [4] I. Mihaylov, M. Nisheva, and D. Vassilev, "Application of machine learning models for survival prognosis in breast cancer studies," *Information (Switzerland)*, vol. 10, no. 3, 2019, doi: 10.3390/INFO10030093.
- [5] A. Sharma and P. K. Mishra, "Performance analysis of machine learning based optimized feature selection approaches for breast cancer diagnosis," *International Journal of Information Technology (Singapore)*, vol. 14, no. 4, pp. 1949–1960, Jun. 2022, doi: 10.1007/S41870-021-00671-5/METRICS.
- [6] S. S. Yadav and S. M. Jadhav, "Thermal infrared imaging based breast cancer diagnosis using machine learning techniques," *Multimed Tools Appl.*, vol. 81, no. 10, pp. 13139–13157, Apr. 2022, doi: 10.1007/S11042-020-09600-3/METRICS.
- [7] Z. Zeng et al., "Identifying Breast Cancer Distant Recurrences from Electronic Health Records Using Machine Learning," *J Healthc Inform Res.*, vol. 3, no. 3, pp. 283–299, Sep. 2019, doi: 10.1007/S41666-019-00046-3/METRICS.
- [8] A. Alzu'bi, H. Najadat, W. Doulat, O. Al-Shari, and L. Zhou, "Predicting the recurrence of breast cancer using machine learning algorithms," *Multimed Tools Appl.*, vol. 80, no. 9, pp. 13787–13800, Apr. 2021, doi: 10.1007/S11042-020-10448-W/METRICS.
- [9] S. Rani, M. Kaur, and M. Kumar, "Recommender system: prediction/diagnosis of breast cancer using hybrid machine learning algorithm," *Multimed Tools Appl.*, vol. 81, no. 7, pp. 9939–9948, Mar. 2022, doi: 10.1007/S11042-022-12144-3/METRICS.
- [10] N. Arya and S. Saha, "Multi-modal advanced deep learning architectures for breast cancer survival prediction," *Knowl Based Syst.*, vol. 221, p. 106965, Jun. 2021, doi: 10.1016/J.KNOSYS.2021.106965.
- [11] S. Lei et al., "Global patterns of breast cancer incidence and mortality: A population - based cancer registry data analysis from 2000 to 2020," *Cancer Commun.*, vol. 41, no. 11, p. 1183, Nov. 2021, doi: 10.1002/CAC2.12207.
- [12] OMS, "Breast cancer." <https://www.who.int/es/news-room/fact-sheets/detail/breast-cancer> (accessed Jan. 02, 2023).
- [13] P. Xuan, L. Jia, T. Zhang, N. Sheng, X. Li, and J. Li, "LDAPred: A method based on information flow propagation and a convolutional neural network for the prediction of disease-associated lncRNAs," *Int J Mol Sci.*, vol. 20, no. 18, Sep. 2019, doi: 10.3390/IJMS20184458.
- [14] A. Turcu-Stiolica et al., "Diagnostic Accuracy of Machine-Learning Models on Predicting Chemo-Brain in Breast Cancer Survivors Previously Treated with Chemotherapy: A Meta-Analysis," *Int J Environ Res Public Health.*, vol. 19, no. 24, Dec. 2022, doi: 10.3390/IJERPH192416832.
- [15] O. Iparraguirre-Villanueva et al., "The Public Health Contribution of Sentiment Analysis of Monkeypox Tweets to Detect Polarities Using the CNN-LSTM Model," *Vaccines* 2023, Vol. 11, Page 312, vol. 11, no. 2, p. 312, Jan. 2023, doi: 10.3390/VACCINES11020312.
- [16] M. F. Aslan, "A hybrid end-to-end learning approach for breast cancer diagnosis: convolutional recurrent network," *Computers and Electrical Engineering.*, vol. 105, p. 108562, Jan. 2023, doi: 10.1016/J.COMPELECENG.2022.108562.
- [17] P. Wang et al., "Cross-task extreme learning machine for breast cancer image classification with deep convolutional features," *Biomed Signal Process Control.*, vol. 57, p. 101789, Mar. 2020, doi: 10.1016/J.BSPC.2019.101789.
- [18] Y. Kaya and F. Kuncan, "A hybrid model for classification of medical data set based on factor analysis and extreme learning machine: FA + ELM," *Biomed Signal Process Control.*, vol. 78, p. 104023, Sep. 2022, doi: 10.1016/J.BSPC.2022.104023.
- [19] A. Ahuja, L. Al-Zogbi, and A. Krieger, "Application of noise-reduction techniques to machine learning algorithms for breast cancer tumor identification," *Comput Biol Med.*, vol. 135, p. 104576, Aug. 2021, doi: 10.1016/J.COMPBIOMED.2021.104576.
- [20] O. Iparraguirre-Villanueva et al., "Search and classify topics in a corpus of text using the latent dirichlet allocation model," *Indonesian Journal of Electrical Engineering and Computer Science.*, vol. 30, no. 1, pp. 246–256, Apr. 2023, doi: 10.11591/IJEECS.V30.I1.PP246-256.
- [21] A. Ahuja, L. Al-Zogbi, and A. Krieger, "Application of noise-reduction techniques to machine learning algorithms for breast cancer tumor identification," *Comput Biol Med.*, vol. 135, p. 104576, Aug. 2021, doi: 10.1016/J.COMPBIOMED.2021.104576.
- [22] K. N. Chitrala, M. Nagarkatti, P. Nagarkatti, and S. Yeguvapalli, "Analysis of the TP53 deleterious single nucleotide polymorphisms impact on estrogen receptor alpha-p53 interaction: A machine learning approach," *Int J Mol Sci.*, vol. 20, no. 12, Jun. 2019, doi: 10.3390/IJMS20122962.
- [23] H. Y. Tsai et al., "Integration of Clinical and CT-Based Radiomic Features for Pretreatment Prediction of Pathologic Complete Response to Neoadjuvant Systemic Therapy in Breast Cancer," *Cancers (Basel).*, vol. 14, no. 24, Dec. 2022, doi: 10.3390/CANCERS14246261.

- [24] F. Xiong, X. Cao, X. Shi, Z. Long, Y. Liu, and M. Lei, "A machine learning-Based model to predict early death among bone metastatic breast cancer patients: A large cohort of 16,189 patients," *Front Cell Dev Biol*, vol. 10, Dec. 2022, doi: 10.3389/FCCELL.2022.1059597.
- [25] S. Khozama and A. M. Mayya, "A New Range-based Breast Cancer Prediction Model Using the Bayes' Theorem and Ensemble Learning," *Information Technology and Control*, vol. 51, no. 4, pp. 757–770, Dec. 2022, doi: 10.5755/J01.ITC.51.4.31347.
- [26] E. Taghizadeh, S. Heydarheydari, A. Saberi, S. JafarpourNesheli, and S. M. Rezaei, "Breast cancer prediction with transcriptome profiling using feature selection and machine learning methods," *BMC Bioinformatics*, vol. 23, no. 1, Dec. 2022, doi: 10.1186/S12859-022-04965-8.
- [27] R. R. Kadhim and M. Y. Kamil, "Comparison of machine learning models for breast cancer diagnosis," *IAES International Journal of Artificial Intelligence*, vol. 12, no. 1, pp. 415–421, Mar. 2023, doi: 10.11591/IJAI.V12.II.PP415-421.
- [28] S. A. Mortazavi, "Machine Learning Models for Predicting Breast Cancer Risk in Women Exposed to Blue Light from Digital Screens," *J Biomed Phys Eng*, Apr. 2022, doi: 10.31661/JBPE.V010.2105-1341.
- [29] A. Lemay et al., "Improving the repeatability of deep learning models with Monte Carlo dropout," Feb. 2022, doi: 10.1038/S41746-022-00709-3.
- [30] D. Clement, E. Agu, J. Obayemi, S. Adeshina, and W. Soboyejo, "Breast Cancer Tumor Classification Using a Bag of Deep Multi-Resolution Convolutional Features," *Informatics*, vol. 9, no. 4, p. 91, Oct. 2022, doi: 10.3390/INFORMATICS9040091.
- [31] H. Wang, Y. Li, S. A. Khan, and Y. Luo, "Prediction of breast cancer distant recurrence using natural language processing and knowledge-guided convolutional neural network," *Artif Intell Med*, vol. 110, p. 101977, Nov. 2020, doi: 10.1016/J.ARTMED.2020.101977.
- [32] S. M. and J. Joy, "A machine learning based framework for assisting pathologists in grading and counting of breast cancer cells," *ICT Express*, vol. 7, no. 4, pp. 440–444, Dec. 2021, doi: 10.1016/J.ICTE.2021.02.005.
- [33] S. Chattopadhyay, A. Dey, P. K. Singh, D. Oliva, E. Cuevas, and R. Sarkar, "MTRRE-Net: A deep learning model for detection of breast cancer from histopathological images," *Comput Biol Med*, vol. 150, p. 106155, Nov. 2022, doi: 10.1016/J.COMPBIOMED.2022.106155.
- [34] Y. Zhou, Y. Niu, Q. Luo, and M. Jiang, "Teaching learning-based whale optimization algorithm for multi-layer perceptron neural network training," *Mathematical Biosciences and Engineering*, vol. 17, no. 5, pp. 5987–6025, Sep. 2020, doi: 10.3934/MBE.2020319.
- [35] O. Iparraguirre-Villanueva et al., "Convolutional Neural Networks with Transfer Learning for Pneumonia Detection," *IJACSA) International Journal of Advanced Computer Science and Applications*, vol. 13, no. 9, p. 2022, Accessed: Jan. 04, 2023. [Online]. Available: [www.ijacsa.thesai.org](http://www.ijacsa.thesai.org).
- [36] Fix Evelyn; Hodges Joseph, "Discriminatory Analysis. Nonparametric," 1951.
- [37] P. A. Jaskowiak and R. J. G. B. Campello, "Comparing Correlation Coefficients as Dissimilarity Measures for Cancer Classification in Gene Expression Data", Accessed: Nov. 07, 2022. [Online]. Available: <https://www.researchgate.net/publication/260333185>.
- [38] F. Nigsch, A. Bender, B. van Buuren, J. Tissen, E. Nigsch, and J. B. O. Mitchell, "Melting point prediction employing k-nearest neighbor algorithms and genetic parameter optimization," *J Chem Inf Model*, vol. 46, no. 6, pp. 2412–2422, 2006, doi: 10.1021/CI060149F/SUPPL\_FILE/CI060149F-FILE002.XLS.
- [39] P. Kasemsumran and E. Boonchieng, "EEG- Based Motor Imagery Classification Using Novel String Grammar Fuzzy K-Nearest Neighbor Techniques with One Prototype in Each of Classes," 2020 International Conference on Artificial Intelligence in Information and Communication, ICAIIC 2020, pp. 742–745, Feb. 2020, doi: 10.1109/ICAIIIC48513.2020.9065236.
- [40] R. Saini, "Integrating Vegetation Indices and Spectral Features for Vegetation Mapping from Multispectral Satellite Imagery Using AdaBoost and Random Forest Machine Learning Classifiers," *Geomatics and Environmental Engineering*, vol. 17, no. 1, pp. 57–74, 2023, doi: 10.7494/GEOM.2023.17.1.57.
- [41] T. H. Nguyen and A. T. Vu, "An Efficient Differential Evolution for Truss Sizing Optimization Using AdaBoost Classifier," *CMES - Computer Modeling in Engineering and Sciences*, vol. 134, no. 1, pp. 429–458, 2023, doi: 10.32604/CMES.2022.020819.
- [42] N. H. A. Malek, W. F. W. Yaacob, Y. B. Wah, S. A. Md Nasir, N. Shaadan, and S. W. Indratno, "Comparison of ensemble hybrid sampling with bagging and boosting machine learning approach for imbalanced data," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 29, no. 1, p. 598, Jan. 2022, doi: 10.11591/IJEECS.V29.II.PP598-608.
- [43] A. Almomani et al., "Age and Gender Classification Using Backpropagation and 焔 agging 焔 lgorithms," *Computers, Materials & Continua*, vol. 74, no. 2, pp. 3045–3062, 2023, doi: 10.32604/CMC.2023.030567.
- [44] M. Fan, K. Xiao, L. Sun, S. Zhang, and Y. Xu, "Automated Hyperparameter Optimization of Gradient Boosting Decision Tree Approach for Gold Mineral Prospectivity Mapping in the Xiong'ershan Area," *Minerals*, vol. 12, no. 12, Dec. 2022, doi: 10.3390/MIN12121621.
- [45] S. Priya, N. K. Karthikeyan, and D. Palanikkumar, "Pre Screening of Cervical Cancer Through Gradient Boosting Ensemble Learning Method," *Intelligent Automation and Soft Computing*, vol. 35, no. 3, pp. 2673–2685, 2023, doi: 10.32604/IASC.2023.028599.
- [46] S. Das, M. S. Imtiaz, N. H. Neom, N. Siddique, and H. Wang, "A hybrid approach for Bangla sign language recognition using deep transfer learning model with random forest classifier," *Expert Syst Appl*, vol. 213, Mar. 2023, doi: 10.1016/J.ESWA.2022.118914.
- [47] M. Imane, C. Rahmoune, M. Zair, and D. Benazzouz, "Bearing fault detection under time-varying speed based on empirical wavelet transform, cultural clan-based optimization algorithm, and random forest classifier," *JVC/Journal of Vibration and Control*, Jan. 2021, doi: 10.1177/10775463211047034.

# Placement of Edge Servers in Mobile Cloud Computing using Artificial Bee Colony Algorithm

Bing Zhou\*, Bei Lu, Zhigang Zhang

College of Artificial Intelligence, Jiaozuo University, Jiaozuo, Henan, 454000, China

**Abstract**—Utilizing smart mobile devices for entertainment, education, and social networking has grown recently. Even though mobile applications are getting more sophisticated and resource-intensive, the computing power of mobile devices is still constrained. Mobile phone applications can perform better by shifting parts of their functions to cloud servers. However, because the cloud is frequently positioned far from mobile phone users, there may be a significant and unpredictable delay in the data transfer between users and the cloud. It is crucial for mobile applications since customers value rapid responses greatly. Users of mobile phones can get close-up access to information technology and cloud computing services thanks to mobile edge computing. In this article, the main goal is to use an artificial bee colony meta-innovative algorithm to solve the problem of placing edge servers in mobile edge computing. Moreover, load balancing between servers is one of the challenges discussed in this article. To deal with this issue, determination the locations of the servers using considering the distribution of workload between servers as a cost function in the artificial bee colony algorithm is a focused issue in this study. The results of the proposed method are compared with the load balancing criteria. The results of K-means compared to the clustering method show the superiority of the proposed method with regard to the loading criteria compared to this clustering method.

**Keywords**—Artificial bee colony algorithm, k-means, server placement, mobile cloud computing

## I. INTRODUCTION

Now-a-days, smart mobile devices have become more critical for smart life [1-3]. In order to accelerate the growth of smart cities, it is necessary to increase the data transfer rate and the amount of infrastructure that is utilized by the services and application programs that make up the smart city [4, 5]. Due to the limited resources of mobile devices, mobile phone users do not experience the same levels of satisfaction as desktop device users, despite the fact that mobile applications are developing and becoming more computationally intensive [6]. Principally, we have a number of users in this issue. Users have a number of requests, and servers have the duty to respond to users' requests. Since each user has a specific position, each server must respond to the requests of users who are close to that server as much as possible. User requests are distributed between servers as a load. Loading part of the activities of mobile phone applications in remote-rich clouds is an effective technique to boost their performance [6–10]. However, because the cloud is frequently situated very distant from mobile phone users, there may be a significant and unpredictable delay in the data transfer between the two parties. It is undesirable for mobile applications, including augmented reality apps and

mobile multiplayer gaming systems, where consumers value quick responses highly.

Mobile phone edge computing enables users of mobile phones to access cloud computing and information technology services inside radio access network coverage [11,12]. In order to decrease the delay, this technique moves the capacity for calculations and storage from the main network to the edge network. Edge servers can be installed nearby so that devices can offload part of the duties associated with their mobile applications, thus enhancing the quality of mobile user experiences.

Most studies have concentrated on offloading mobile user workloads to cloudlets to reduce mobile device energy consumption, and this strategy implies that clouds are already in place [13-16]. The impact of putting edge servers and loading mobile phone user workload onto them on mobile application performance has not received much consideration. In mobile edge computing, edge server deployment should take into account both the ideal location and the maximum number of edge servers. For this purpose, three points should be considered, which include the physical location of the edge servers, the planning of the number of edge servers, and the network deployment requirements. According to these three cases, in this article, we intend to choose the optimal location for deploying edge servers by using an artificial bee colony meta-initiative algorithm. The placement of the edge server in mobile edge computing environments has not received much research attention.

Numerous studies on cloudlet location have been conducted recently [17, 18]. The typical definition of a *cloud* is a network of computers that operate as a loader for users of mobile phones [19, 20]. Users of mobile phones can access edge servers close to base stations in the context of mobile edge computing [21]. To reduce the access delay between mobile phone users and distant clouds, edge servers can currently be considered download locations for mobile users. By sending calculations and storage capacity from the primary network to the Edge server, this problem is resolved [22]. It is believed that there are many similarities between cloudlet placement and edge servers [23, 24]. Mike Jia et al. [25] presented a load shedding system model for multi-user mobile task loading. They studied cloudlet placement and mobile user allocation to cloudlet. Then, with the aim of activating the location of the cloudlet in areas with high user density, they implemented an algorithm and, while balancing their workload, assigned mobile phone users to the contracted packages.

\*Corresponding Author.

Zichuan Zhu and colleagues [16] utilized numerous wireless programs to investigate the cloudlet positioning issue. In order to reduce the access time between mobile users and cloudlets servicing users, they first defined this problem as a new cloudlet placement problem that assigns  $K$  locations to various key crucial sites. Second, they created an effective solution after defining the issue as integer linear programming. Studies [26, 27] about the computing burden for mobile edge cloud computing are also available. In a multi-channel wireless interference environment, Hu Chen et al. [28] investigated the issue of multi-user computing load for mobile edge cloud computing. They created a distributed computing offloading technique that attained the Nash equilibrium and characterized the distributed computing load decision problem for mobile phone users as a multi-user computing offloading game. Mike Jia et al. [25] only focused on the workload balance of cloudlets. Other studies [29-31] only on delay Access are examined, despite the fact that the aforementioned studies are successful in addressing the issue of cloudlet placement.

In this study, it is researched to find the position of servers that result in the optimization of coverage and balanced response to users' requests using the artificial bee colony meta-innovative method, which is motivated by the inspirations of previous studies. The steps of this study include studying the methods of the past, introducing the proposed method, and simulating this method. For this purpose, the MATLAB programming environment is used.

In this environment, the initial conditions for deploying edge servers in cloud computing are simulated using an artificial data set. In this simulation, variables such as the location of servers, base stations, and the number of user requests are considered. The results of this simulation are compared with the method based on K-means clustering. For the implementation of the Internet of Things, mobile edge computing has grown in importance [32]. Edge servers are smaller and less vulnerable to attacks since they are scattered over the mobile edge computing ecosystem. In addition, edge servers may function as private clouds to help with the issue of data leaking. There seems to be a need to employ a way to locate the server based on mobile edge computing technologies.

The rest of this paper presents the proposed method in Section II. The results and discussion discuss in Section III. The paper concludes in Section IV.

## II. PROPOSED METHOD

In this section, a method for determining the location of servers in mobile edge computing is introduced. This method uses an artificial bee colony meta-initiative algorithm. In this method, by using this algorithm, we try to optimize a cost function, which leads to determining the location of the servers in such a way that the load balance between the servers is observed.

The primary objective of the problem of locating servers for mobile edge computing is to locate servers for proper and ideal user request response. The quantity of users must be taken into account first for this. These users each have a number of requests, which they submit to the server. The

servers must respond to the requests of the users. In fact, each server receives a number of requests from users and must respond to these requests. According to the stated conditions, the location of the servers should be such that the servers follow the main goal. The main purpose of servers, in addition to responding to users' requests, is to distribute the load proportionally and optimally among themselves. In fact, servers should be located, so user requests are distributed among them. The position of the users and the number of requests of the users are known in this problem. However, the position of the servers should be answered as unknown in this problem. An artificial bee colony meta-innovative algorithm has been used to determine the location of the servers. The steps of the proposed method include:

defining and assigning values to the used variables,

creating a data set,

simulating the proposed method on the data set to determine the location of the servers, and

simulating the K-means clustering method on the data set to determine the location of the servers.

The comparison of the results obtained from the proposed method and the K-means clustering method with the load balancing criterion is introduced in the following stages of the proposed method.

### A. Variables Under Study

In order to determine the status of the servers in the problem under consideration, the users and the requests of the users must be determined. According to the situation and the number of user requests, the values of the servers should be determined as unknown. In this study, the number of users, the position of users in the form of  $x$  and  $y$  coordinates in two-dimensional space, and the number of user requests are considered variables. Since real data is unavailable in this problem, synthetic data is used to value the studied variables. As an example, the position of the number of users is shown in Fig. 1.

The values of the users' coordinates are randomly assigned in the range  $[a-0]$ . The number of requests of each user is also a random number in the range of  $[1-w]$ . By assigning values to these variables, the used data set is produced.

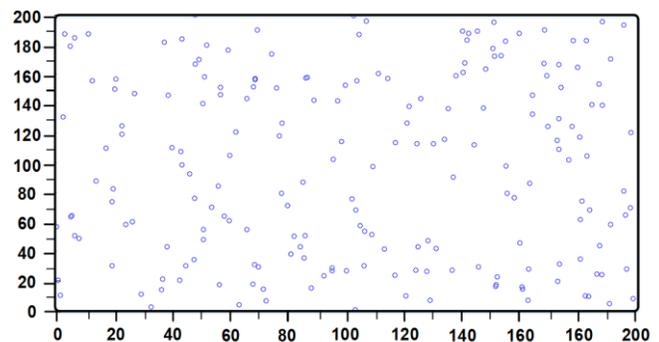


Fig. 1. The position of users in a specific area.

### B. Artificial Bee Colony Algorithm to Servers Determine Location

In the issue of determining the location of the servers, it is necessary to determine the location of each server so that the servers can answer the requests of the users in a way that the division and distribution of the load are balanced between them. To determine load balance, it is necessary to have a relationship that determines the amount of load between servers, calculated according to Eq. (1). The artificial bee colony algorithm determines the optimal value for the load balance relationship [33].

$$WB = \sqrt{\frac{\sum_{i=1}^k (T^i - T)^2}{k}} \quad (1)$$

Where  $k$  represents the number of servers,  $T^i$  represents the number of requests that server  $i$  should respond.  $T$  indicates the average number of requests of all servers. In this regard, the lower the WB value indicates the existence of balance and load balance between servers. In order to respond to each user's request, a server must be assigned, and that server will answer the user's request. The location of servers and users are shown with two-dimensional coordinates. The closest server answers each user's request. Euclidean distance relationship is used to determine the distance between servers and a user. If we have  $k$  servers to respond to the request of user  $j$  according to equation (2), server  $i$  should be selected so that these servers have the smallest distance from user  $j$ .

$$\arg \min_{i \in (1,k)} (\sqrt{(x_j - x_i)^2 + (y_j - y_i)^2}) \quad (2)$$

Where  $x_i$  and  $y_i$  represent the x and y position of server  $i$ ,  $x_j$  and  $y_j$  represent the x and y position of server  $j$ . Since the goal and the unknown problem is to determine the location of the servers, by using the artificial bee colony algorithm, the initial location of the servers is generated as the initial population. By going through the steps of the bee colony algorithm to minimize equation (1) as a cost function, the most optimal position is determined for the servers. The stages of the artificial bee colony algorithm are according to Fig. 2.

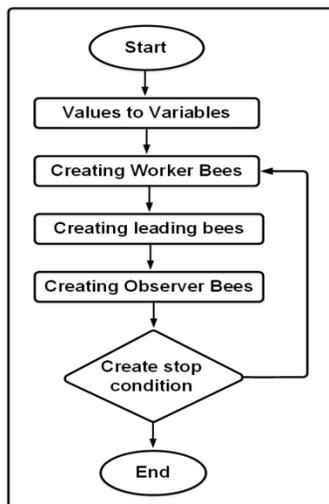


Fig. 2. Steps of artificial bee colony algorithm.

By having the location of server  $k$ , the closest server to that user is determined for each user using equation (2). After the closest server has been assigned to each user, the load balance between the servers can be determined by equation (2). For each population produced in the steps of the artificial bee colony algorithm, the fitness of this population is also determined by equation (2). The stages of production of worker bees, observers, and leaders lead to searching in the state space and producing better answers to the investigated problem.

### C. Proposed Method Evaluation

In the problem of determining the location of the servers in calculating the mobile edge, two important criteria and metrics for the location of the servers are considered. The combination of these two criteria, which includes the determination of the closest server to each user and the distribution of user requests, leads to the creation of a standard called load balance. Since the servers have to perform the users' tasks in the end, the servers' location should be such that all servers can be balanced in terms of workload. In the problem of determining the location of servers, load balancing means balanced distribution of user requests between servers. The proposed method uses the load balance criterion as a cost function in the artificial bee colony algorithm. The results obtained from the proposed method have been compared with the K-means clustering method to evaluate the proposed method. In the K-means clustering method, users are placed in  $k$  clusters, and each cluster's center is considered the server's location. In fact, the K-means clustering algorithm groups two-dimensional points that are the coordinates of users into  $k$  clusters. Users in a cluster send their requests to the server head of the current cluster. With this assumption, the load balance is also determined by the K-means clustering method according to equation (1). The evaluation of the proposed and K-means clustering methods is done by comparing the load balance value. The method with less load balance is the more suitable method for determining the location of the servers.

## III. RESULT AND DISCUSSION

In this section, the simulation of the proposed method is examined. According to the proposed method stated in the previous section, the steps of the proposed method are simulated in this section. This section includes the introduction of the data set, performing some tests on the data set, and reviewing the results obtained from the various tests.

### A. Dataset Used

To simulate the proposed method, it is necessary to set and select the variables and data set. The first part of the data set is the simulation environment. In the simulation environment, the position of users and servers is determined. The simulation environment is a square with side  $a$ . The second part of the data set is the users. Variables related to users include the number of users, the position of each user, and the number of requests for each user. The third part of the data set is the servers. The variables related to servers also include the number of servers and the location of each server. The required variables are introduced in Table I.

The position of each user is shown as a point in two-dimensional coordinates. The value of  $x$  for the  $i$ -th user is in a

specific range in the interval between  $xf-xe$ . The value of  $y$  for the  $i$ -th user is also within the specified range in the interval  $yf-ye$ . Each server also has a location. The position of each server, like the position of each user, is shown as a point in two-dimensional space. The range of  $x$  and  $y$  values for each server is the same as the range of these values for users. Each user also has a number of requests. The number of requests of user  $i$  is determined by a number in the range  $wf-we$ . Since users and servers are shown with two-dimensional coordinates, the distance between a user  $i$  and server  $j$  is represented by the symbol  $d_{ij}$ . Euclidean distance, according to Eq. (1) is used to calculate the distance between a user and a server.

The position of each user and the number of requests of each user are randomly determined within the specified range. The position of each server is also unknown and is calculated and determined by the proposed method. In the proposed method, an artificial bee colony meta-initiative algorithm is used. The variables of this algorithm are the number of repetitions and the number of the initial population. The values of these variables are shown in Table II.

### B. Results Evaluation

In this section, the variables stated in the previous section are set. Having the data set in this section, several different tests are performed. In these experiments, the proposed K-means clustering methods are compared with each other. These different experiments which are simulating and determining the position of the servers. The main difference is in the valuing of the studied variables. This section consists of the results of three experiments. In these tests, the number of users, the position of users, and the range determined for the position of users are considered different from each other. The values of the variables in the first experiment are according to Table III.

The results obtained from this experiment with four different executions are according to Table IV. This experiment has been done with four different implementations.

The position of users and servers in each of these four tests in the proposed method and the K-means clustering method is according to Fig. 3 to 10.

The values of the cost function in the artificial bee colony algorithm are according to Fig. 11 to 14.

In the first experiment with the criterion of workload balance, the results of the proposed method are better than the k-means clustering method. In this test, even in Fig. 7, the selection of two servers with a close location shows that the proposed method has considered even close locations in the search space. Although this happens randomly, the ability to search in the state space in the proposed method is better than the k-means method. In these four implementations, the proximity of the servers to each other or the distance between the servers did not have such a big effect on the results of the load balance. In other words, the artificial bee colony method in different repetitions is able to find several answers in different situations, but with the results of the load balance approximately to find the same. In contrast, in the k-means method, the distance between the servers in different executions is not so different from each other. Although the answers generated for the location of the servers are different,

the distance between the servers in each of these four executions is greater than the executions of the proposed method. In the graphs of the cost function values in the proposed method, the proposed method has converged in many iterations. The values of the variables in the second experiment are according to Table V.

The results obtained from this experiment with four different executions are according to Table VI.

Each of these four tests in the proposed method and the k-means clustering method is according to Fig. 15 to 26.

In the second experiment, with the criterion of workload balance, the results of the proposed method are better than the k-means clustering method. In this test, the number of users and requests are the same as in the first test, but the position of the users is randomly assigned in a larger range than in the first test. The results of the load balance values in this experiment are almost similar to the values in the first experiment since the number of user requests did not differ from the first experiment. In this case, it is only proportional to the situation of new users of the servers. The proposed method and the k-means method determine that the servers are positioned so that the load balance values are not much different from the first test. The position of the servers in the proposed method is better than the k-means method in terms of distance variation in each execution. In fact, the proposed method has produced answers that the results of the location of the servers are close to each other in some executions and far from each other in others. Despite the production of various answers, the results of load balance in different implementations with the proposed method are not much different. The values of the variables in the third experiment are according to Table VII.

The results obtained from this experiment with four different executions are according to Table VIII.

Each of these four tests in the proposed method and the k-means clustering method is according to Fig. 27 to 34.

The values of the cost function in the artificial bee colony algorithm are according to Fig. 35 to 38.

In the third experiment with the criterion of workload balance, the results of the proposed method are better than the k-means clustering method. In this experiment, the number of users, the position of the users, and the number of user requests differ from the first experiment. The number of users in this test is twice the number in the first test. The position that can be selected for assigning values to the coordinates of users is also twice the first test. The number of requests of 200 users in the third test is similar to the first. That is, the number of requests of the first 100 users in the third test is similar to the 100 users of the first test. The number of requests of the second 100 users is similar to the 100 users of the first test. In this experiment, the location of the servers in the proposed method is better than the k-means method in terms of distance variation in each execution. In fact, the proposed method has produced answers that the results of the location of the servers are close to each other in some executions and far from each other in others. Despite the production of various answers, the results of load balance in different executions with the proposed method are not much different. In this section, the main goal is to

simulate the proposed method. The results of the proposed method have been compared with the k-means clustering method with the criterion of load balance. In order to simulate the proposed method, three tests have been performed in this

section. The results obtained from these tests show the superiority of the proposed method over the k-means clustering method with load balancing criteria and generating various answers for the servers' location.

TABLE I. VARIABLES USED

Variable Name	Variable Symbol
Limited simulation environment	A square with side a
Number of users	n
The position of each user i	$x_i \in [x_f - x_e]$ $y_i \in [y_f - y_e]$
Number of requests per user i	$w_i \in [y_f - y_e]$
Number of servers	k
Server position j	$x_j \in [x_f - x_e]$ $y_j \in [y_f - y_e]$
The distance between user i and server j	$d_{ij}$

TABLE II. THE VALUES OF THE VARIABLES OF AN ARTIFICIAL BEE COLONY META-INITIATIVE ALGORITHM

Artificial bee colony algorithm variable name	Value
Number of repetitions	50
Primary population	10

TABLE III. THE VALUES OF THE VARIABLES IN THE FIRST EXPERIMENT

Variable Name	Variable Symbol
Limited simulation environment	A square with side 200
Number of users	100
The position of each user i	$x_i \in [0-200]$ $y_i \in [0-200]$
Number of requests per user i	$w_i \in [1-5]$
Number of servers	5

TABLE IV. THE RESULTS OF THE FIRST EXPERIMENT

Tests Name	$WB_{acb}$	$WB_{kmeans}$
The first test, the first performance	3.6	13.3026
The first test of the second performance	4.0694	19.3742
The first test of the third performance	4.6648	8.8408
The first test of the fourth performance	3.763	9.0863

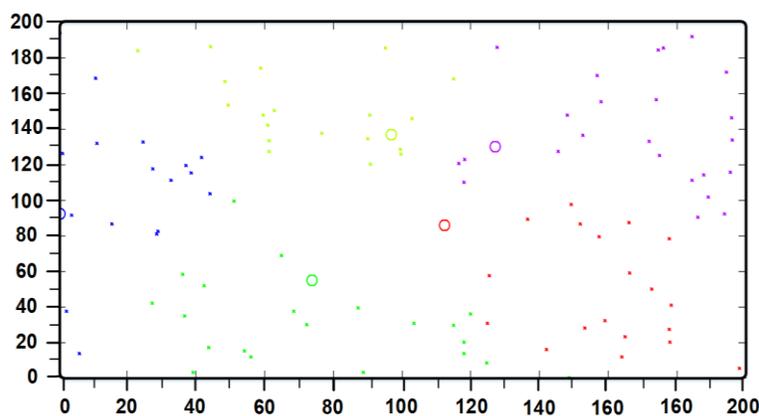


Fig. 3. The position of the servers in the first test of the first execution with the proposed method.

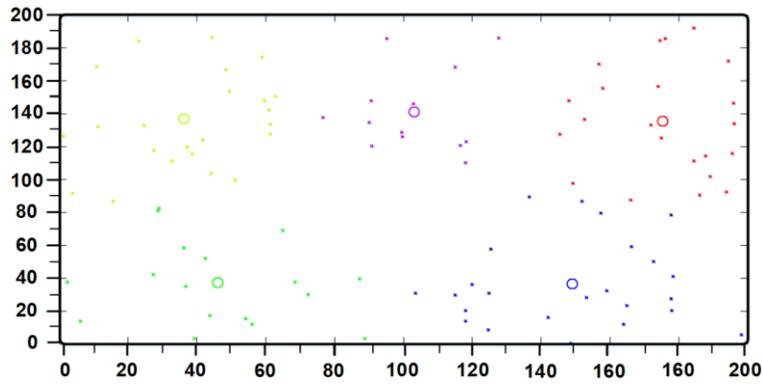


Fig. 4. The position of the servers in the first test of the first run with the k-means method.

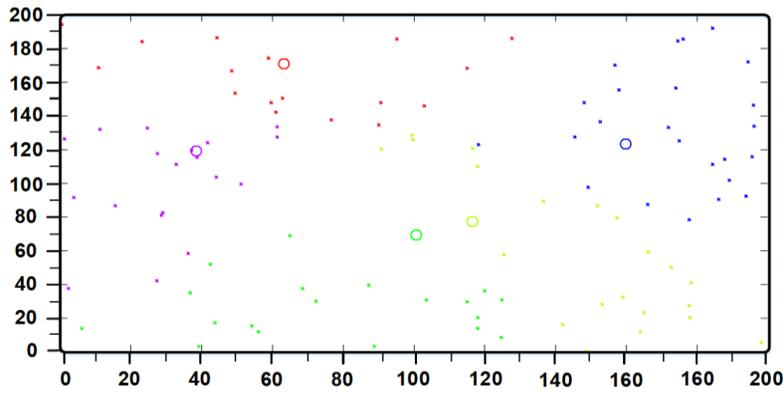


Fig. 5. The position of the servers in the first test of the second execution with the proposed method.

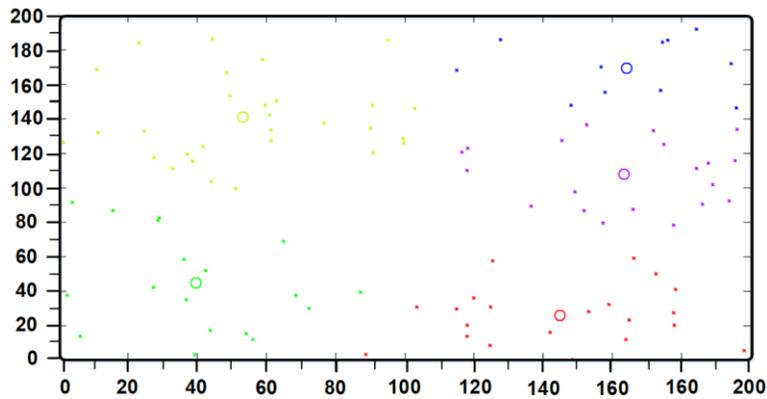


Fig. 6. The position of the servers in the first test of the second run with the k-means method.

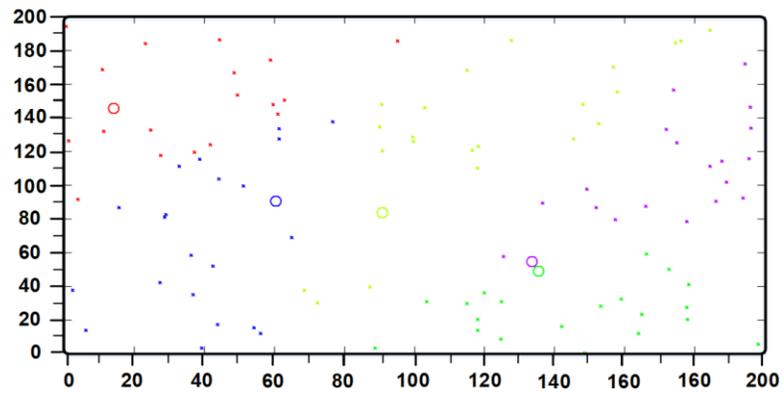


Fig. 7. The position of the servers in the first test of the third execution with the proposed method.

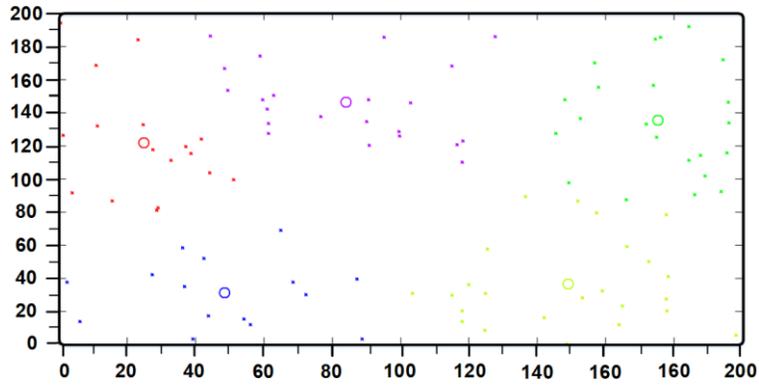


Fig. 8. The position of the servers in the first test of the third run with the k-means method.

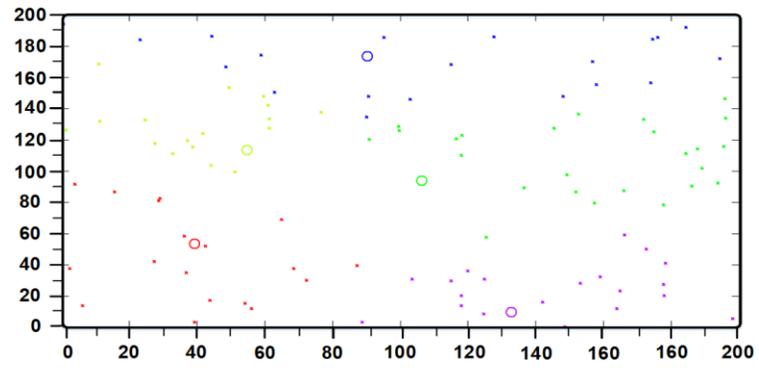


Fig. 9. The position of the servers in the first test of the fourth implementation with the proposed method.

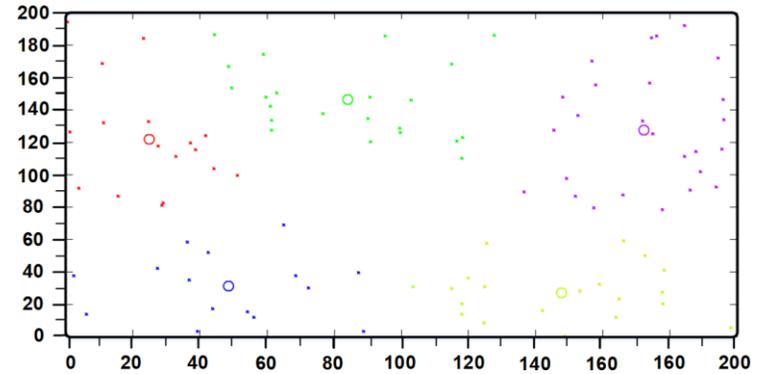


Fig. 10. The position of the servers in the first test of the fourth run with the k-means method.

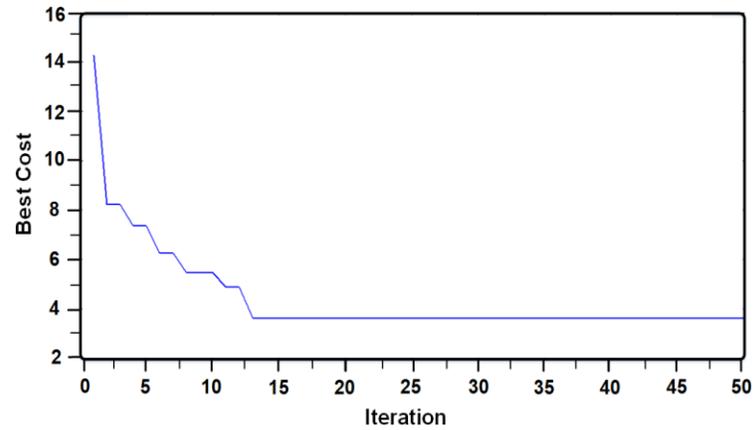


Fig. 11. The value of the cost function in the first experiment of the first execution with the proposed method.

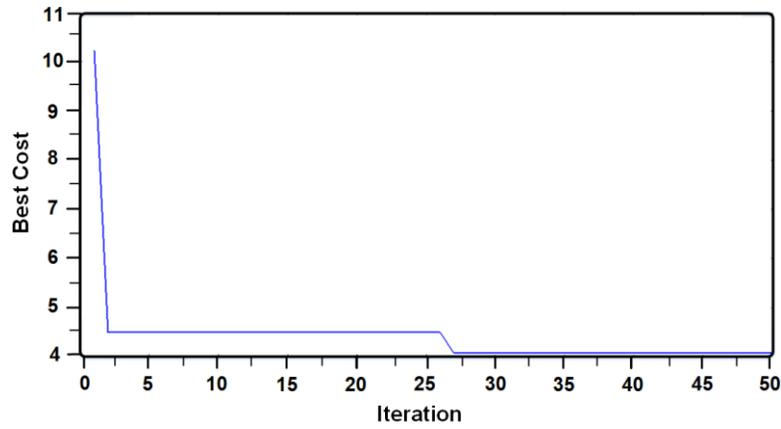


Fig. 12. The value of the cost function in the first experiment of the second implementation with the proposed method.

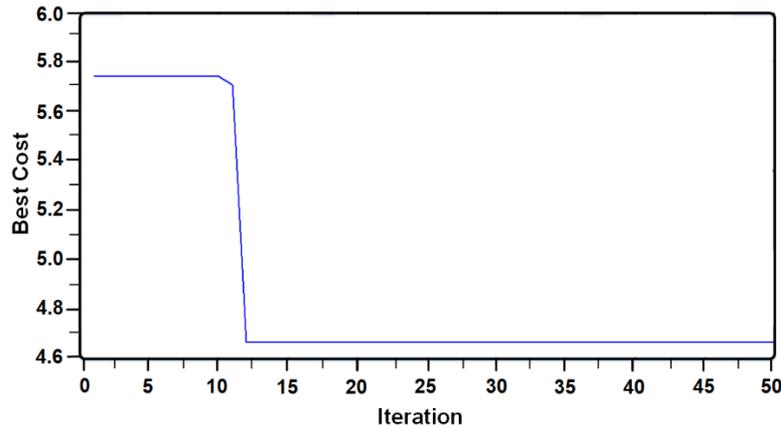


Fig. 13. The value of the cost function in the first experiment of the third implementation with the proposed method.

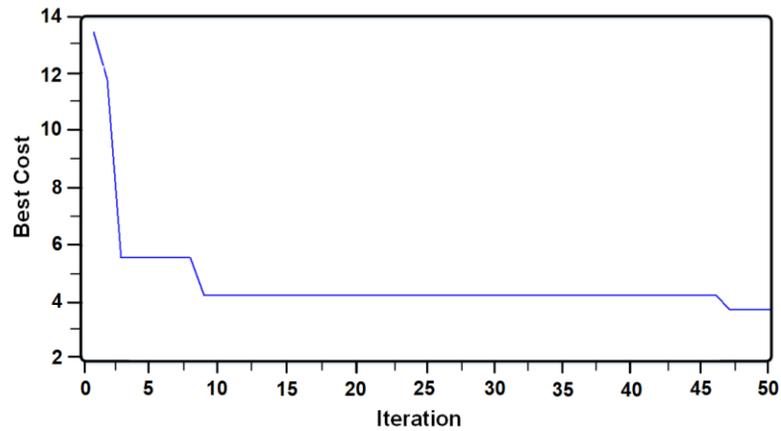


Fig. 14. The value of the cost function in the first experiment of the fourth implementation with the proposed method.

TABLE V. THE VALUES OF THE VARIABLES IN THE SECOND EXPERIMENT

Variable Name	Variable Symbol
Limited simulation environment	A square with side 500
Number of users	100
The position of each user $i$	$x_i \in [0-500]$ $y_i \in [0-500]$
Number of requests per user $i$	$w_i \in [1-5]$
Number of servers	5

TABLE VI. THE RESULTS OF THE SECOND EXPERIMENT

Tests Name	$WB_{acb}$	$WB_{kmeans}$
The second test, the first performance	3.9699	10.4957
The second test of the second performance	4.2615	10.4957
The second test of the third performance	3.0594	11.6516
The second test of the third performance	4.9558	10.4957

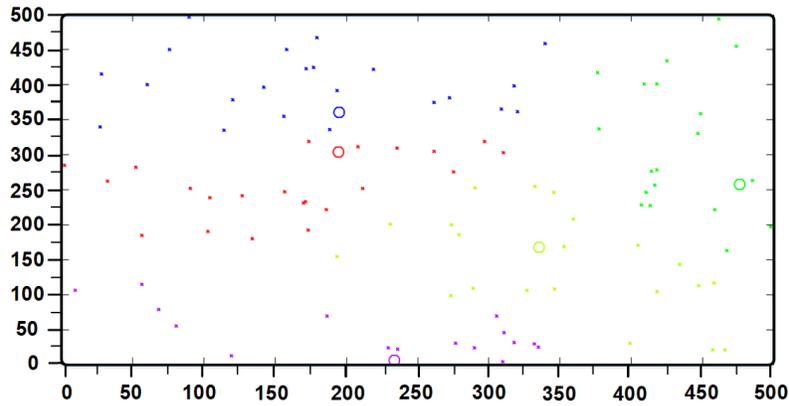


Fig. 15. The position of the servers in the second test of the first execution with the proposed method.

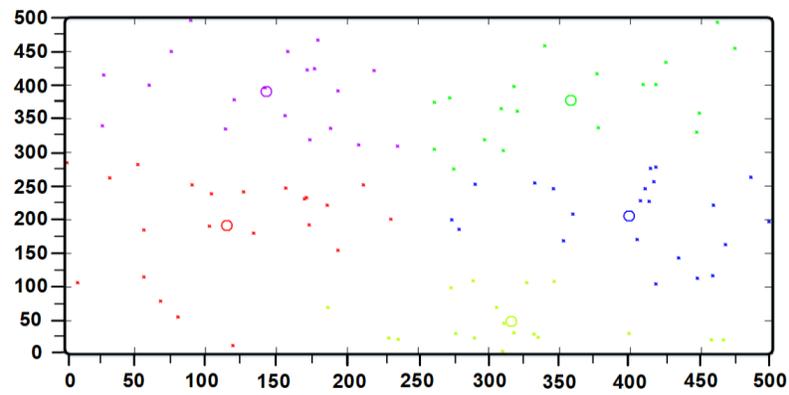


Fig. 16. The position of the servers in the second test of the first run with the k-means method.

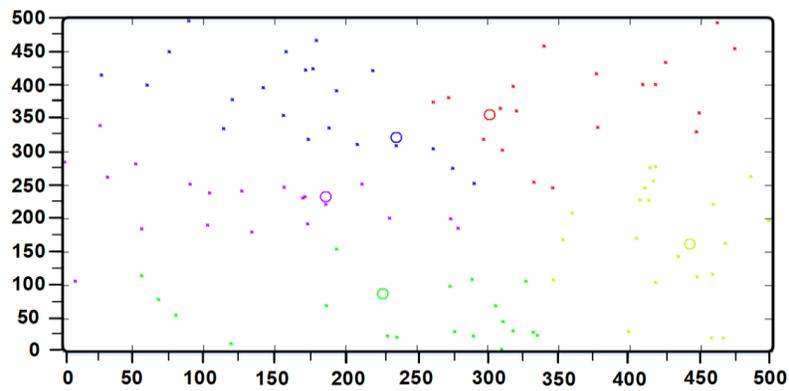


Fig. 17. The position of the servers in the second test of the second execution with the proposed method.

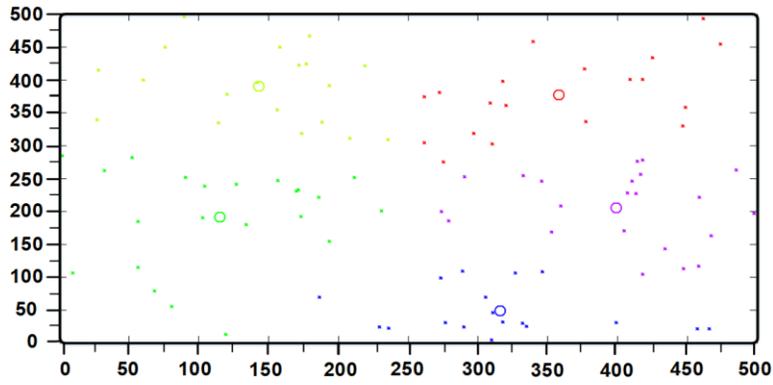


Fig. 18. The position of the servers in the second test of the second run with the k-means method.

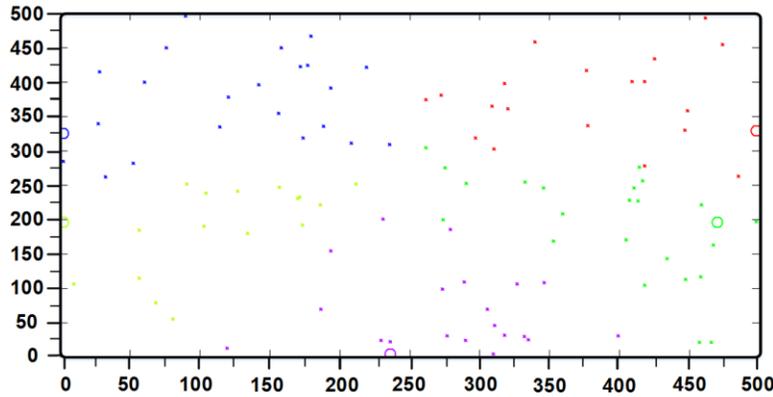


Fig. 19. The position of the servers in the second test of the third execution with the proposed method.

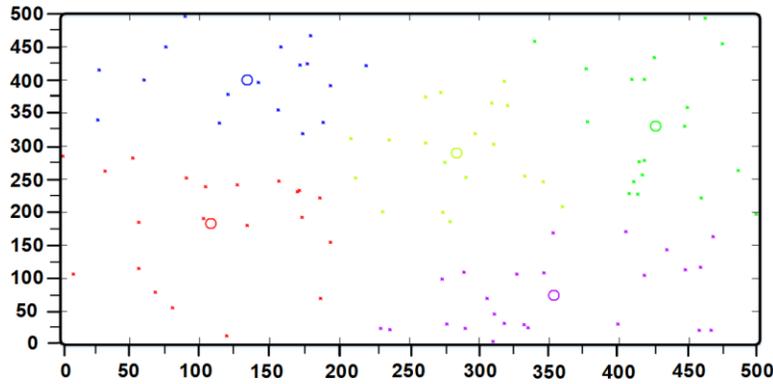


Fig. 20. The position of the servers in the second test of the third run with the k-means method.

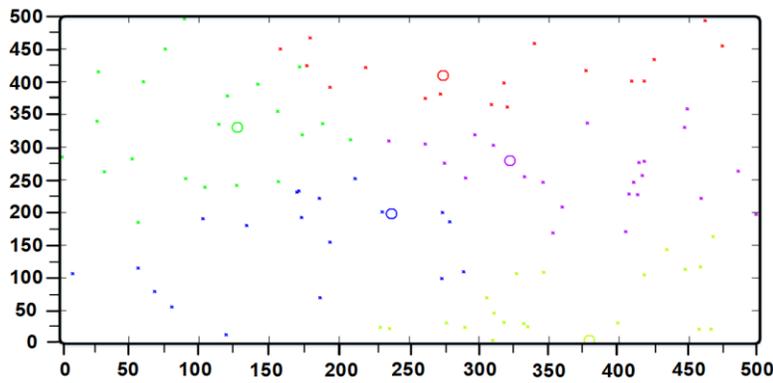


Fig. 21. The position of the servers in the second test of the fourth implementation with the proposed method.

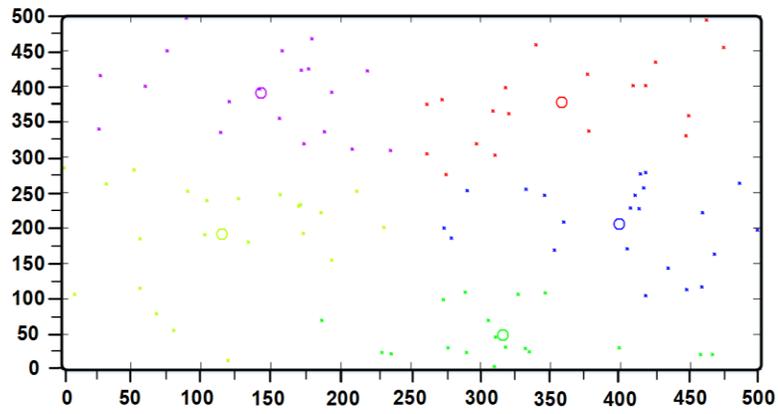


Fig. 22. The position of the servers in the second test of the fourth run with the k-means method.

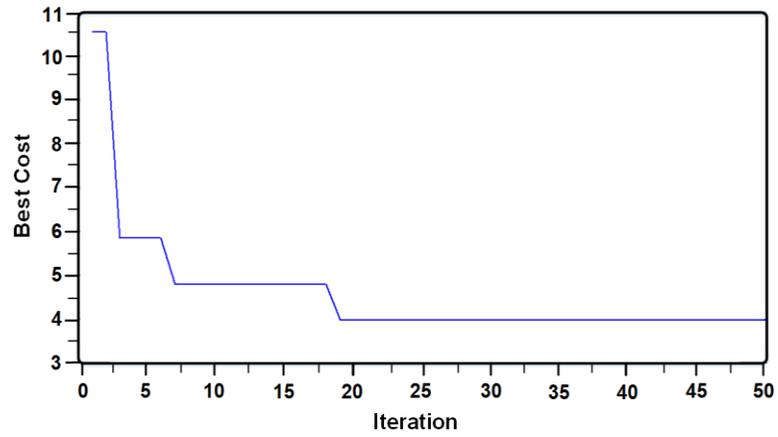


Fig. 23. The value of the cost function in the second experiment of the first execution with the proposed method.

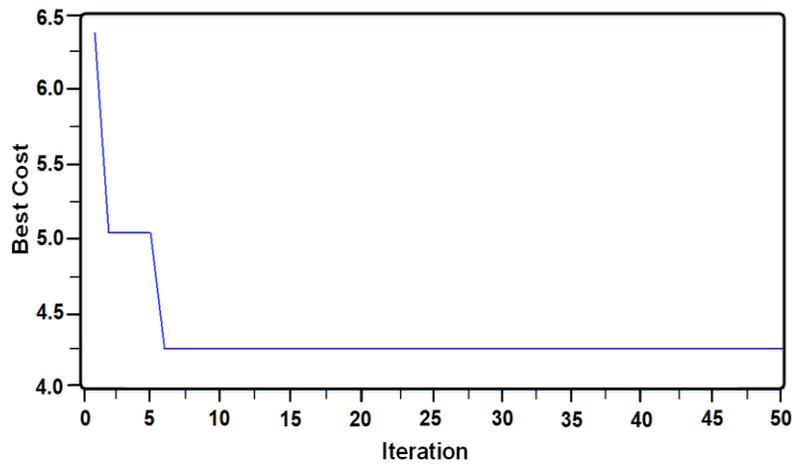


Fig. 24. The value of the cost function in the second experiment of the second implementation with the proposed method.

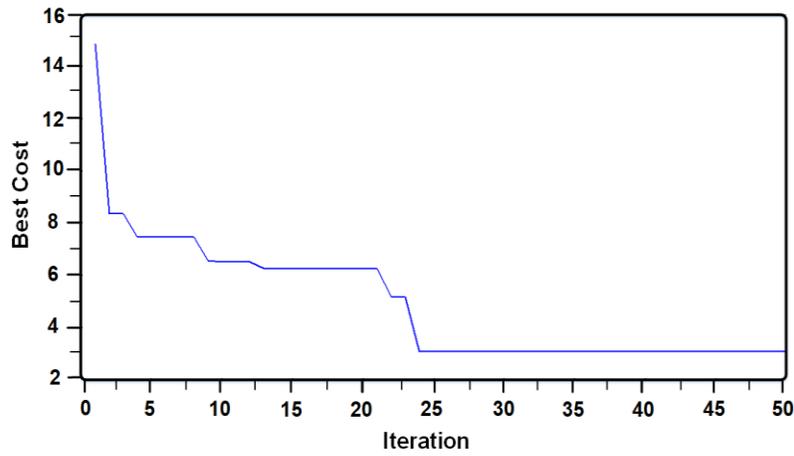


Fig. 25. The value of the cost function in the second experiment of the third implementation with the proposed method.

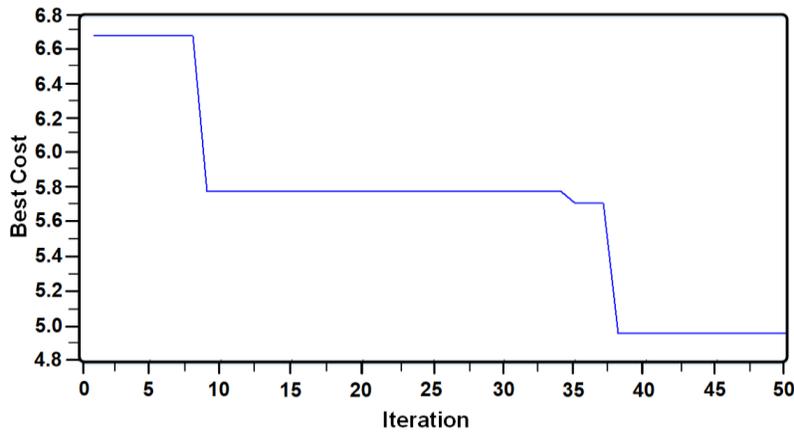


Fig. 26. The value of the cost function in the second experiment of the fourth implementation with the proposed method

TABLE VII. THE VALUES OF THE VARIABLES IN THE THIRD EXPERIMENT

Variable Name	Variable Symbol
Limited simulation environment	A square with side 200
Number of users	200
The position of each user $i$	$x_i \in [0-200]$ $y_i \in [0-200]$
Number of requests per user $i$	The first 100 users are the same as the first 100 test users, the second 100 users are the same as the first 100 test users
Number of servers	5

TABLE VIII. THE RESULTS OF THE THIRD EXPERIMENT

Tests Name	$WB_{acb}$	$WB_{kmeans}$
The third test, the first performance	6.7705	25.9892
The third test of the second performance	10.1902	25.6562
The third test of the third performance	5.2383	22.6945
The third test of the third performance	8.1633	22.2764

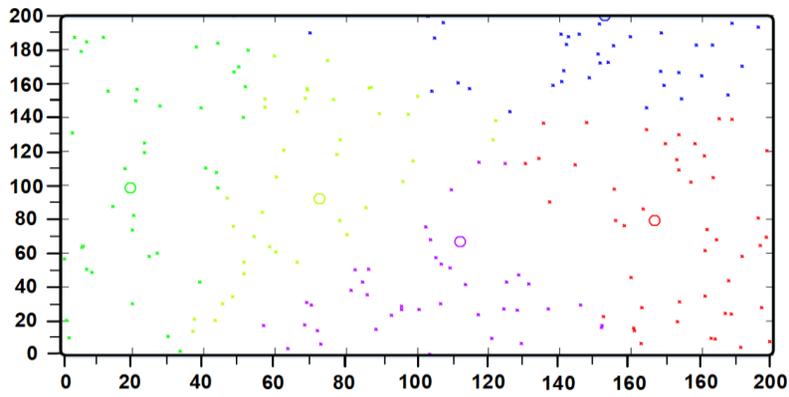


Fig. 27. The position of the servers in the third test of the first execution with the proposed method.

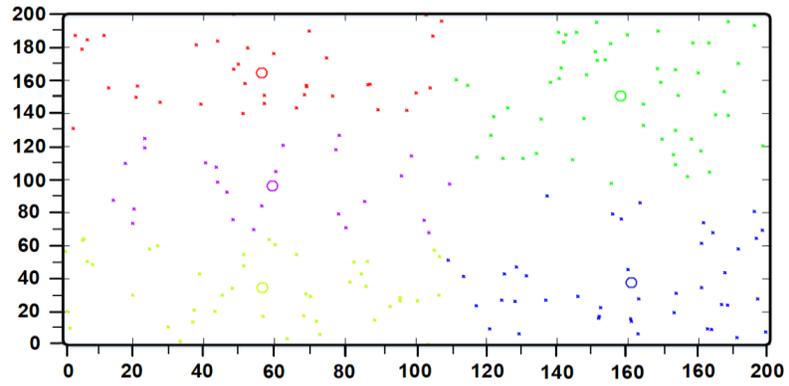


Fig. 28. The position of the servers in the third test of the first run with the k-means method.

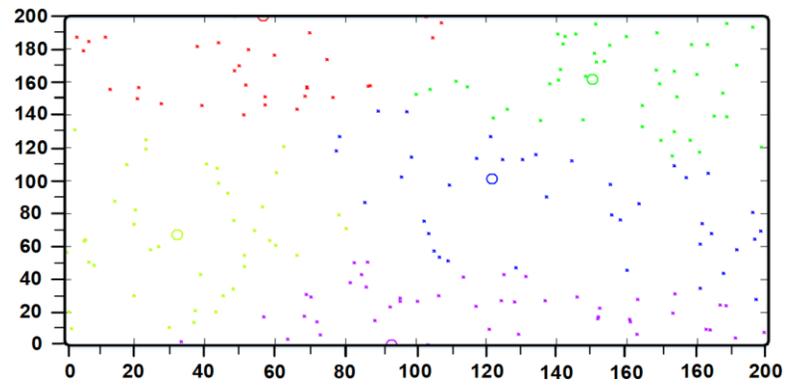


Fig. 29. The position of the servers in the third test of the second execution with the proposed method.

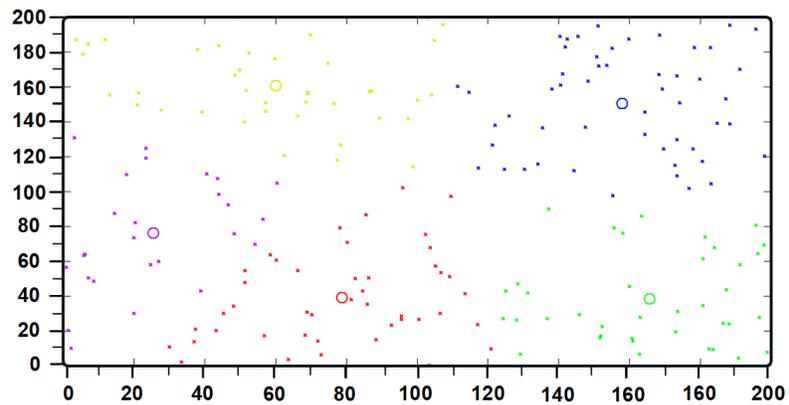


Fig. 30. The position of the servers in the third test of the second run with the k-means method.

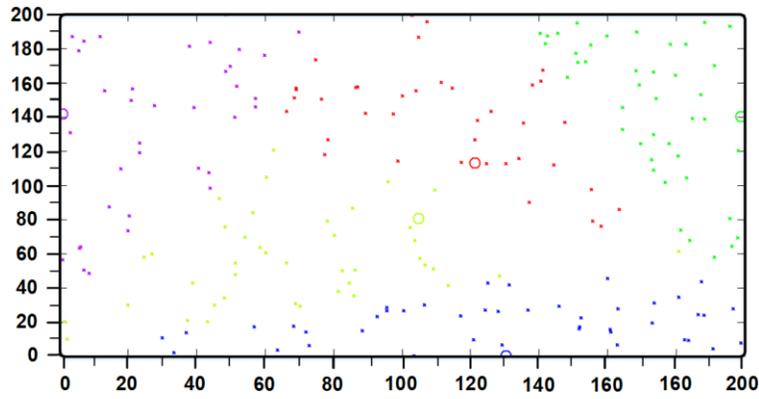


Fig. 31. The position of the servers in the third test of the third execution with the proposed method.

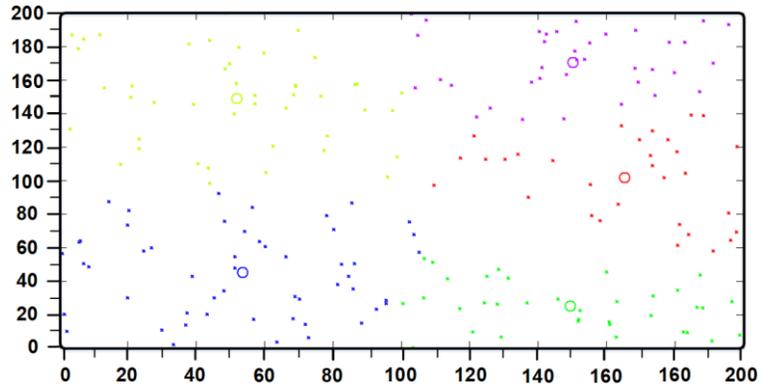


Fig. 32. The position of the servers in the third test of the third run with the k-means method.

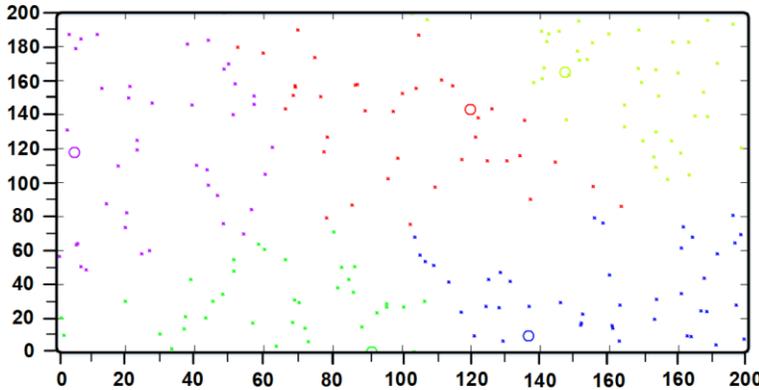


Fig. 33. The position of the servers in the third test of the fourth implementation with the proposed method.

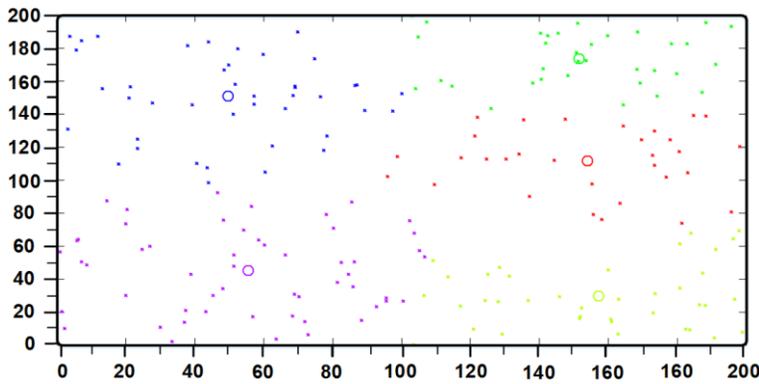


Fig. 34. The position of the servers in the third test of the fourth run with the k-means method.

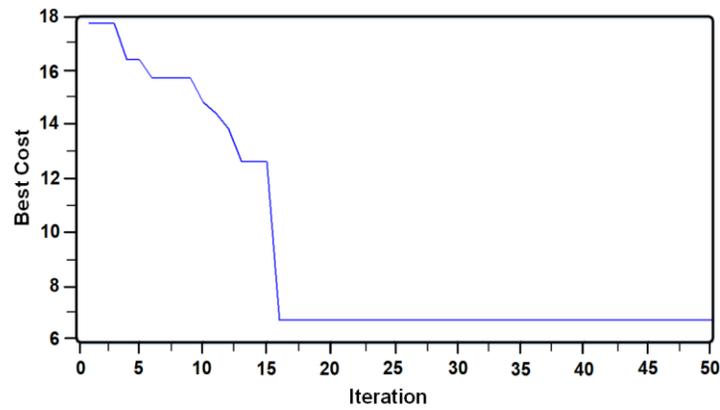


Fig. 35. The value of the cost function in the third experiment of the first execution with the proposed method.

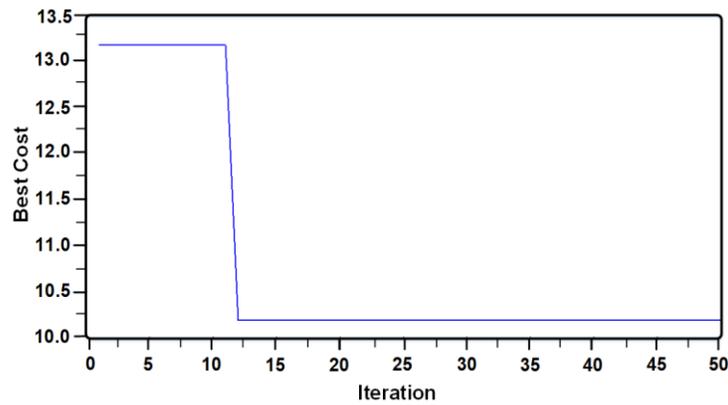


Fig. 36. The value of the cost function in the third experiment of the second implementation with the proposed method.

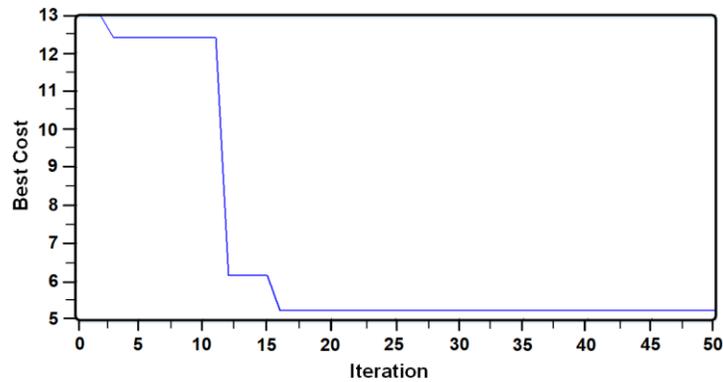


Fig. 37. The value of the cost function in the third experiment of the third implementation with the proposed method.

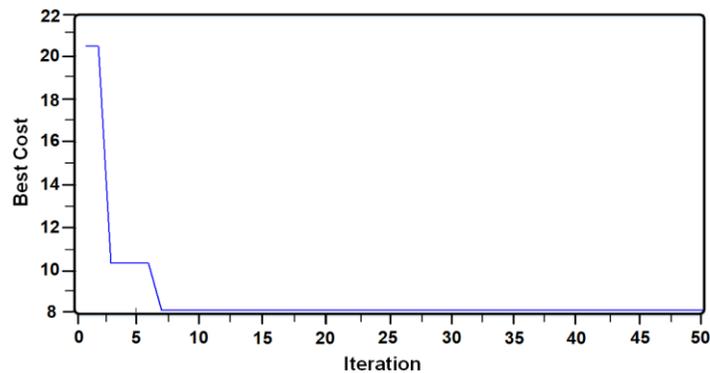


Fig. 38. The value of the cost function in the third experiment of the fourth implementation with the proposed method.

#### IV. CONCLUSION

In the problem of determining the location of the servers, it should be possible to determine the location of the servers in the edge computing environment in mobile computing in such a way that by having  $k$  servers, these servers respond to the requests of the users in such a way that the load of the servers is balanced. In other words, the servers should be in a position to be able to respond to the requests of nearby users, and the number of requests that the servers respond to should be balanced. A criterion called load is used to balance and distribute user requests between servers. The major contribution of this study is to employ a meta-innovative artificial bee colony algorithm to address the issue of where to locate edge servers for mobile edge computing. Moreover, one of the difficulties covered in this essay is load balancing between servers. This study's main focus is on determining the server placements using the artificial bee colony method while taking workload distribution between servers into account as a cost function. The proposed method's outcomes are contrasted with the load balancing criterion. The comparison of K-means results to the clustering approach demonstrates the proposed method presented results superiority with regard to the loading criteria. For future study, the proposed method can be implemented in real scenarios. Furthermore, other optimization algorithms including Particle Swarm Optimization (PSO) can be investigated and the result compared to current study to address the better solution.

#### ACKNOWLEDGMENTS

Training plan for young backbone teachers in Henan Province (No.2018GGJS267).

#### REFERENCES

- [1] Dinh, H. T., Lee, C., Niyato, D., & Wang, P. (2013). A survey of mobile cloud computing: architecture, applications, and approaches. *Wireless communications and mobile computing*, 13(18), PP 1587-1611.
- [2] Taleb, T., Dutta, S., Ksentini, A., Iqbal, M., & Flinck, H. (2017). Mobile edge computing potential in making cities smarter. *IEEE Communications Magazine*, 55(3), PP 38-43.
- [3] Sabella, D., Vaillant, A., Kuure, P., Rauschenbach, U., & Giust, F. (2016). Mobile-edge computing architecture: The role of MEC in the Internet of Things. *IEEE Consumer Electronics Magazine*, 5(4), PP 84-91.
- [4] Kosta, S., Aucinas, A., Hui, P., Mortier, R., & Zhang, X. (2012, March). Thinkair: Dynamic resource allocation and parallel execution in the cloud for mobile code offloading. In *2012 Proceedings IEEE Infocom* (pp. 945-953). IEEE.
- [5] Kemp, R., Palmer, N., Kielmann, T., & Bal, H. (2010, October). Cuckoo: a computation offloading framework for smartphones. In *International Conference on Mobile Computing, Applications, and Services* (pp. 59-79). Springer, Berlin, Heidelberg.
- [6] Chun, B. G., Ihm, S., Maniatis, P., Naik, M., & Patti, A. (2011, April). Clonecloud: elastic execution between mobile device and cloud. In *Proceedings of the sixth conference on Computer systems* (pp. 301-314).
- [7] Ahmed, E., Gani, A., Sookhak, M., Ab Hamid, S. H., & Xia, F. (2015). Application optimization in mobile cloud computing: Motivation, taxonomies, and open challenges. *Journal of Network and Computer Applications*, 52, PP 52-68.
- [8] Patel, M., Naughton, B., Chan, C., Sprecher, N., Abeta, S., & Neal, A. (2014). Mobile-edge computing introductory technical white paper. White paper, mobile-edge computing (MEC) industry initiative, 29, PP 854-864.
- [9] Cuervo, E., Balasubramanian, A., Cho, D. K., Wolman, A., Saroiu, S., Chandra, R., & Bahl, P. (2010, June). Maui: making smartphones last longer with code offload. In *Proceedings of the 8th international conference on Mobile systems, applications, and services* (pp. 49-62).
- [10] Beck, M. T., Feld, S., Fichtner, A., Linnhoff-Popien, C., & Schimper, T. (2015, February). ME-VoLTE: Network functions for energy-efficient video transcoding at the mobile edge. In *2015 18th International Conference on Intelligence in Next Generation Networks* (pp. 38-44). IEEE.
- [11] Taleb, T., Dutta, S., Ksentini, A., Iqbal, M., & Flinck, H. (2017). Mobile edge computing potential in making cities smarter. *IEEE Communications Magazine*, 55(3), PP 38-43.
- [12] Li, H., Dong, M., Ota, K., & Guo, M. (2016). Pricing and repurchasing for big data processing in multi-clouds. *IEEE Transactions on Emerging Topics in Computing*, 4(2), PP 266-277.
- [13] Ahmed, E., Akhunzada, A., Whaiduzzaman, M., Gani, A., Ab Hamid, S. H., & Buyya, R. (2015). Network-centric performance analysis of runtime application migration in mobile cloud computing. *Simulation Modelling Practice and Theory*, 50, PP 42-56.
- [14] Chun, B. G., & Maniatis, P. (2010, June). Dynamically partitioning applications between weak devices and clouds. In *Proceedings of the 1st ACM Workshop on Mobile Cloud Computing & Services: Social Networks and Beyond* (pp. 1-5).
- [15] Xu, Z., Liang, W., Xu, W., Jia, M., & Guo, S. (2015). Efficient algorithms for capacitated cloudlet placements. *IEEE Transactions on Parallel and Distributed Systems*, 27(10), PP 2866-2880.
- [16] Jia, M., Cao, J., & Liang, W. (2015). Optimal cloudlet placement and user to cloudlet allocation in wireless metropolitan area networks. *IEEE Transactions on Cloud Computing*, 5(4), PP 725-737.
- [17] Magurawalage, C. M. S., Yang, K., Hu, L., & Zhang, J. (2014). Energy-efficient and network-aware offloading algorithm for mobile cloud computing. *Computer Networks*, 74, PP 22-33.
- [18] Yao, H., Bai, C., Xiong, M., Zeng, D., & Fu, Z. (2017). Heterogeneous cloudlet deployment and user-cloudlet association toward cost effective fog computing. *Concurrency and Computation: Practice and Experience*, 29(16), e3975.
- [19] Andrews, J. G., Buzzi, S., Choi, W., Hanly, S. V., Lozano, A., Soong, A. C., & Zhang, J. C. (2014). What will 5G be?. *IEEE Journal on selected areas in communications*, 32(6), PP 1065-1082.
- [20] Morgado, A., Huq, K. M. S., Mumtaz, S., & Rodriguez, J. (2018). A survey of 5G technologies: regulatory, standardization and industrial perspectives. *Digital Communications and Networks*, 4(2), PP 87-97.
- [21] Wang, T., Zhou, J., Zhang, G., Wei, T., & Hu, S. (2019). Customer perceived value-and risk-aware multiserver configuration for profit maximization. *IEEE Transactions on Parallel and Distributed Systems*, 31(5), PP 1074-1088.
- [22] Liu, Y., Peng, M., Shou, G., Chen, Y., & Chen, S. (2020). Toward edge intelligence: Multiaccess edge computing for 5G and Internet of Things. *IEEE Internet of Things Journal*, 7(8), PP 6722-6747.
- [23] Chen, M., Guo, S., Liu, K., Liao, X., & Xiao, B. (2020). Robust computation offloading and resource scheduling in cloudlet-based mobile cloud computing. *IEEE Transactions on Mobile Computing*, 20(5), PP 2025-2040.
- [24] Chen, M., Guo, S., Liu, K., Liao, X., & Xiao, B. (2020). Robust computation offloading and resource scheduling in cloudlet-based mobile cloud computing. *IEEE Transactions on Mobile Computing*, 20(5), PP 2025-2040.
- [25] Xiang, H., Xu, X., Zheng, H., Li, S., Wu, T., Dou, W., & Yu, S. (2016, December). An adaptive cloudlet placement method for mobile applications over GPS big data. In *2016 IEEE Global Communications Conference (GLOBECOM)* (pp. 1-6). IEEE.
- [26] Clinch, S., Harkes, J., Friday, A., Davies, N., & Satyanarayanan, M. (2012, March). How close is close enough? Understanding the role of cloudlets in supporting display appropriation by mobile users. In *2012 IEEE international conference on pervasive computing and communications* (pp. 122-127). IEEE.
- [27] Liu, J., Ahmed, E., Shiraz, M., Gani, A., Buyya, R., & Qureshi, A. (2015). Application partitioning algorithms in mobile cloud computing:

- Taxonomy, review and future directions. *Journal of Network and Computer Applications*, 48, PP 99-117.
- [28] Gu, F., Niu, J., Qi, Z., & Atiqzaman, M. (2018). Partitioning and offloading in smart mobile devices for mobile cloud computing: State of the art and future directions. *Journal of Network and Computer Applications*, 119, PP 83-96.
- [29] Satyanarayanan, M., Bahl, P., Caceres, R., & Davies, N. (2009). The case for vm-based cloudlets in mobile computing. *IEEE pervasive Computing*, 8(4), PP 14-23.
- [30] Wolbach, A., Harkes, J., Chellappa, S., & Satyanarayanan, M. (2008, June). Transient customization of mobile computing infrastructure. In *Proceedings of the First Workshop on Virtualization in Mobile Computing* (pp. 37-41).
- [31] Tao, M., Ota, K., & Dong, M. (2017). Foud: Integrating fog and cloud for 5G-enabled V2G networks. *IEEE Network*, 31(2), PP 8-13.
- [32] Kosta, S., Aucinas, A., Hui, P., Mortier, R., & Zhang, X. (2012, March). Thinkair: Dynamic resource allocation and parallel execution in the cloud for mobile code offloading. In *2012 Proceedings IEEE Infocom* (pp. 945-953). IEEE.
- [33] Wang, S., Zhao, Y., Xu, J., Yuan, J., & Hsu, C. H. (2019). Edge server placement in mobile edge computing. *Journal of Parallel and Distributed Computing*, 127, PP 160-168.

# Erythematous-Squamous Disease Detection using Best Optimized Estimators of ANN

Rajashekar Deva<sup>1</sup>, Dr G .Narsimha<sup>2</sup>

Asst.Prof. Department of C.S.E., Methodist College of Engineering &Technology, Abids, Hyderabad, Telangana, 500001, India<sup>1</sup>  
Principal and Professor, JNTUH, Sulthanpur, Sangareddy, Telangana, India<sup>2</sup>

**Abstract**—Medical area focused on automating skin cancer detection after the pandemic era of "Monkey Pox". Previous works proposed ANN mechanisms to classify the type of skin cancer. However, all those models implement layers of ANN with standard estimator components like hidden layers implemented using the ReLu activation function, several neurons are generally a power of two and others, but these values are not always perfect. Few researchers implemented optimization techniques for tuning the estimators of A.I. algorithms, but all those mechanisms require more resources and don't guarantee the best values for each estimator. The proposed method analyzes all the essential estimators of every possible neural network layer. Then it applies a modified version of Bayesian optimization because it avoids the disadvantages of Grid and Random optimization techniques. It picks the best estimator by using the conditional probability of naive Bayesian for every combination.

**Keywords**—Conditional probability; naive Bayesian; bayesian optimization; grid search; optimization techniques; estimators

## I. INTRODUCTION

Most researchers utilize machine learning to identify skin cancers but suffer from overfitting and more resource consumption. The problem can be solved using artificial neural networks, but traditional networks become complicated for smaller datasets. So, the proposed research aims to customize the parameters of the network by performing "Hyper Parameterization".

Hyperparameter tuning is required for behavioural control of the machine-learning model [11]. Hyperparameters can be configured differently for each machine-learning model. Our expected model parameters will yield less-than-ideal outcomes if the hyperparameters are not appropriately tweaked to optimize the loss function [12]. This implies that our model has further problems. When building a model from a particular dataset, hyper-tuning identifies the potential optimal sets of hyperparameters. A single training task executes several trials for hyperparameter tweaking. The task of hyperparameter tweaking involves meta-optimization [27]. The hyperparameter tuner produces the hyperparameter setting, those results in the best-performing model after assessing various hyperparameter settings. By employing the algorithm and the defined limits of hyperparameters, hyperparameter tuning runs multiple training sessions on your dataset to determine which model version is the best.

### A. Hyper Parameter Tuning

The accuracy of models for machine learning can be significantly increased by using hyperparameter adjustment.

Hyperparameter tuning is identifying a collection of appropriate hyperparameter variables for a prediction model and applying this adjusted algorithm to any piece of data. The model's efficiency is maximized by employing that collection of hyperparameters, which minimizes a preset loss function and produces better results with fewer errors. To use an Exhaustive Grid Search in Scikit Learn is a well-known and conventional method for hyperparameter tweaking [13]. Every permutation of each collection of hyperparameters is tested using this procedure. This approach allows us to locate the ideal set of values within the variable search space. Since this approach must test every permutation in the grid size, it typically consumes more computer resources and requires a long time to run. Multiplication of all the variables will determine the size of the parameter grid. Each time a random collection of hyperparameters is tested, the model's performance is recorded. After multiple repetitions, it returns to the mixture with the most significant outcome.

### B. Types of Hyper Tuning

Irrespective of the type of machine and deep learning models, these tuning algorithms help achieve the minimum error rate with minimum learning rate.

1) *Grid search*: Grid search optimization takes much time to compute every possible combination of estimators. Suppose we have two parameters for designing the ANN in which parameter-1 can be estimated in X ways, and parameter-2 can be estimated in Y ways. The system needs X into Y ways to apply grid search for this type of neural network. So this can be claimed as an exhaustive search.

2) *Random search*: Random search only checks some possible estimated values [14]. It randomly selects a few values from each parameter, so this will reduce the number of ways than grid search since random search doesn't cover all possible combinations. So this tuning process only guarantees the best values.

3) *Successive halving*: The S.H.A. (Successive Halving Algorithm) algorithm can optimize hyperparameters and solve multi-armed bandits challenges. The algorithm's primary goal is to correctly determine the best arm within a strict budget, a constrained time or resource [15]. The algorithm consistently evaluates each arrangement. The weakest performers are removed at the conclusion of each round. The procedure repeats itself until only one configuration is left, with the remaining configurations being examined twice as much as in

the previous round. S.H.A. is effective because it uses minimal resources and eliminates data at each level.

Whether or not to look for several configurations in S.H. is still being determined. Some desirable configurations that may initially converge slowly will be eliminated early if  $n$  is large. This makes identifying the best allocation approach in the indefinite time frame possible. S.H. requires dynamic updating. Before using the S.H. technique, the hyperparameters must be manually set. Due to its iterative nature, Successive Halving works well enough on huge datasets. The wide range of cross-validation folds could also be used as the budget for successive Halving. A sufficient budget can prevent good configurations from being terminated too soon. In contrast, an excessive budget can cause subpar configurations to continue for an extended period and waste resources.

4) *Bayesian optimization*: When deciding which set of hyperparameters to examine next, Bayesian optimization takes into consideration previous assessments. It allows itself to concentrate on those regions of the dimensional space that it considers will provide the most hopeful validation scores by selecting its value, also known intelligently. This method often needs fewer iterations to reach the ideal set of hyperparameter values [16]. Most significantly, it ignores those regions of the dimensional space that it thinks won't contribute anything. As a result, only settings predicted to produce a better validation score are sent through for evaluation, reducing the number of repetitions a model must be trained for validation. It is especially helpful when these assessments are at high costs, when derivatives are absent, or the matter in question is non-convex.

The paper is divided into five sections; the introduction discusses the need and types of hyper optimization techniques. The literature survey section discusses the merits and demerits of the existing approaches. The proposed methodology discusses the customization of the neural network layers with the help of enhanced Bayesian optimization. The results and discussion section elaborates on the metrics obtained by the proposed and compares them with the existing ones. The conclusion section discusses the proof of validity by measuring loss and accuracy.

## II. LITERATURE SURVEY

In [1], MehwishDildar et al. offered a thorough analysis of deep learning methods for early skin cancer detection. The paper concentrated on traditional methods for skin cancer diagnosis, such as ANN, K.N.N., CNN, and GAN. The writers created multiple stages of selection standards. This work aims to evaluate existing models and develop the best N.N. technique for skin cancer detection. The data was filtered using an automated search engine that was developed. Neural networks have been taught to categorize photos and differentiate between photographs of various skin disorders.

Additionally compared were SVM, B.P.N., and three-layer N.N. Due to a shortage of different data, artificial neural networks are trained for skin lesions using tiny data sets.

According to the authors, the auto-organization approach, which is still under investigation, may enhance image processing accuracy in the future, particularly in the medical industry.

In [2], Mariam Nawaz et al. developed a deep learning-based segmentation solution that is completely automated. The model uses fuzzy K-means clustering and region-based convolutional neural networks (RCNN). The model applies RCNN using preprocessed data. Using precise localization, faster RCNN could detect skin lesions with accuracy and precision. Various than melanoma, the approach can be applied to other skin conditions. F.K.M. clustering separates the impacted portion of photos from the discovered results. A deep learning framework is faster than RCNN. Faster RCNN relies on the input generic object suggestions, which employ hand-coded models like Edge Box, Selective search, etc. Three datasets are used to apply the model. The proposed model has a chance of being overfitted. This model requires less computing. The F.K.M. model is useful in resolving to overfit.

In [3], Ulzii-OrshikhDorj et al. concentrated on classifying skin cancer using deep CNN and ECOC SVM. The feature extraction method makes use of trained AlexNet. ECOC SVM is employed for classification, along with three fully linked layers. Some photographs in the collection are clear, but others are not because they were pulled from the internet. The work can be expanded by adding the ABCD rule— asymmetry, border, colour, and diameter—for each cancer report. There are four different types of cancer in total. To cut down on noise, the photos are cropped. The pooling layer reduced the size of the input neuron in the CNN. The model's primary goal is quick categorization. Using a deep CNN model, using RGB images allows for detecting actinic keratoses, basal cell carcinoma, melanoma, and squamous cell carcinoma [26].

In [4], Mohammad Ali Kadampur and Sulaiman Al Riyale combined and suggested a model-driven cloud architecture. This is used to build models that help in skin cancer prediction and is based on deep learning algorithms. The metric area beneath the curve for the deep learning algorithm was 99%. Despite the model's lack of programming components, the machine can deal with issues including slowness, accuracy, and a shortage of dermatologists. The suggested model can categorize cell pictures and spot skin cancer. The model is not integrated into the REST API. According to the article, there are now more design options for deep learning classifiers regarding general methods and looping patterns. The paper described the D.L.S. tool's features and showed how to build a deep learning model.

In [5], Andre Esteva et al. used a single CNN-trained end-to-end using images with simple pixels and illness labels as input to demonstrate classification. A computer technique is created that could help doctors and patients keep track of skin blemishes and spot cancer early. Dermatology is built to be automated. No handcrafted components are necessary for the system. By putting the model to the test, the biopsy-proven images are verified. The writers designed the CNN to mimic dermatologists' performance. By putting the biopsy photos to

the test, the model is verified. The authors explored the internal properties that CNN learned using T-SNE. Mobiles can be used to deploy this strategy. Several virtual circumstances can be classified using this method, provided there are enough training instances. Saliency maps are created to see the pixels a network focuses on for prediction.

In [6], Ravi Manne et al. summarised numerous studies on utilizing CNNs to categorize skin lesions. CNN has demonstrated incredible image processing power. The research covered a wide range of adversarial approaches in clinical contexts, including colour balance adjustments and input image rotation and translation that can result in incorrect categorization. The authors also noted the variables influencing the findings. Instead of using all the positive data, which would skew the system and produce only good outcomes, the factors are mentioned as vulnerabilities to adversarial assaults. The inks (blue) in dermoscopic images also hurt CNN classification, suggesting that the model may give accurate results if negative input is included. The authors discovered that improperly categorized photos might produce inaccurate results.

In [7], Khalid M. Hosny et al. introduced a system that automatically classifies skin marks. A deep learning network which is pre-trained serves as the model. AlexNet is trained via data segmentation and fine-tuning. In the system to train and evaluate the suggested model, the ph2 dataset is used. The deep convolutional neural network (DCNN) model divides the three forms of skin cancer visible in colour photographs. Any image can be processed and used in the proposed method. Since it is unnecessary, there is no preprocessing. A softmax layer has taken the place of the categorization on layer. The model eliminates the requirement for photographs with labels. To create a deep neural network, labelled images are necessary. The size of the convolution layer is decreased using the pooling layer. The weights have been refined using backpropagation to include additional weights for categorizing skin lesions. Based on dataset photos, the weights are changed using S.G.D. Stochastic Gradient Descent.

In [8], JinenDaghrir et al. introduced a hybrid technique to diagnose melanoma skin cancer. The model examines any suspect lesions. Three distinct approaches are included in the model. Using a set of features including borders, texture, color, etc, a convolutional neural network and two conventional machine learning classifiers are developed. The ABCDE signals are used in the model as melanoma markers. Based on five features, this trait can distinguish benign skin lesion that develops into malignant melanoma. Using colour enhancement on the RGB's blue component, DoG Filtering is used to detect hair. The 124 x 124-pixel pictures that the CNN architecture that is being suggested uses [24]. The CNN contains nine layers, including three convolutional layers with ReLU activation and three with maximum spatial pooling. Strange skin lesions need to be researched. To achieve better results, adopting semi-supervised learning should be the main focus.

In [9], Vidya M & Dr Maya V Karki suggested an approach that employs the ABCD rule for feature extraction. To extract features for the early diagnosis of skin lesions,

GLCM and H.O.G. are utilized. The preprocessing enhances the clarity and quality of skin lesions while lowering artefacts such as hair, skin color, etc. Using Geodesic Active Contour, which separates each instructional section, segmentation was carried out. For extracting attributes, including symmetry, border, color, and diameter, the ABCD scoring system is employed. The retrieved characteristics are sent straight to the classifiers. A clear skin lesion site can be obtained using dermoscopy, improving visual impact. The obtained accuracy is 97.32%. G.A.C. detects the largest changes to the entire skin lesion, typically produced around its edges. RGB images are transformed into grayscale and put through a median filter to reduce the noise. On the same, neural networks may be used.

TABLE I. MERITS AND FUTURE SCOPE WORK ANALYSIS OF EXISTING SYSTEMS

Author	Algorithm	Merits	Demerits
MehwishDildar	S.L.R.	Suggesting auto organization approaches.	Need to implement automation in the model, focused only on NN-based models.
Mariam Nawaz	Faster RCNN	FKM clustering is also used	Chance to overfit
Ulzii-OrshikhDorj	CNN and ECOC SVM	The noise was removed, and RGB images improved the properties of the image.	Can add ABCD rule and implement to get better results.
Mohammad Ali Kadampur	Deep learning and cloud.	Used D.L.S. tools, can deal with	Not implemented into an API
Andre Esteva	CNN	Uses simple pixels and saliency maps are used.	No required number of training records
Ravi Manne	ESLER	Focussed on many issues like misclassification and dermoscopic images.	Need to include the operating results of various models.
Khalid M. Hosny	DCNN	Transfer learning, any mage can be preprocessed.	The weights keep changing
JinenDaghrir	MCNN	ReLU activation function, two ML classifiers.	Should implement for semi-supervised data.
Vidya M	Hybrid feature extraction	Uses ABCDE signs, images conversion into grayscale	Neural networks provide better results.
Shunichi Jinnai	FRCNN	Robust, high classification accuracy	It would help if you also used N.N. for generalization. The project should be implemented socially.

In [10], Shunichi Jinnai et al. built a dataset by randomly choosing items and annotating them with bounding boxes. The model that is being suggested is a quicker, region-based CNN called FRCNN. By combining convolutional features from R.P.N. and Fast R CNN into a single network, FRCNN is the result—a classification system based on neural networks that use clinical images rather than dermoscopic ones. , Robustness, high classification accuracy, and speed were all displayed by the model. A momentum stochastic gradient

descent optimizer is utilized, with the VGG-16 as a foundation. To employ wearable technology in public, the network must be socially implemented. The skin cancer prognosis should be used to lower treatment expenses. Reduced patient wait times and unneeded visits are desirable [25]. It is important to test the neural network's generalization using images thoroughly. The overall analysis of the existing approaches is presented in Table I.

### III. RESEARCH GAPS IDENTIFIED

- 1) Unsupervised models to solve the non-linear data have increased the dimensionality of the space.
- 2) Traditional neural networks make the model learn more number of generalized features rather than specific elements.
- 3) Backpropagation in neural networks makes the model update the weights more randomly to get optimized results.

### IV. PROPOSED METHODOLOGY

The proposed methodology contains 34 attributes in the dermatology dataset. Using machine learning approaches like decision trees or ensemble methodologies also takes many resources to build the tree for simple datasets. So the proposed model implemented customized neural networks instead of static neural networks. The components of the ANN, along with their estimators, are presented below in Table II.

#### A. Input Layer and its Estimators

In general, the input, without performing any transformations, passes the raw information obtained from the input vector to the hidden layers. It multiplies every value of the feature with random weights and then implements the activation function.

TABLE II. COMPONENT DESCRIPTION OF NEURAL NETWORK

Components	Remarks	Estimators	Description
Input Layer	The input layer, along with the input shape, contains activation and several neurons which can be customized.	Number of neurons	Any integer value. Generally, it is equal to the number of features in the dataset.
		Activation Functions	Based on the differentiable properties of the neurons, they are classified into 10
Hidden Layer	Depending on the activation function applied and other metrics, it transforms the input into the desired result using the dot product.	Kernel Size	This helps the CNN to process the imaging unit by unit because, without this, the system has to process $n*n*3$ at a time. Generally, the system works efficiently with an odd number of filters.
		Padding	The system needs to add some bits in terms of padding values to provide accurate results. In general, two types of paddings are available.
		Stride	It defines the step size, representing the number of pixels to ignore. In general, it will be any n value.
		Activation	Based on the differentiable properties of the neurons, they are classified into 10
Pooling Layer	The neural network size is reduced using the strides and pool size. There are three types of pooling mechanisms.	Pool size	It's a two-dimensional filter that maps the features.
		Stride	It defines the step size, representing the number of pixels to ignore. In general, it will be any n value.
Output Layer	Generally, it is a fully connected layer that produces the desired output based on several neurons and activation functions.	Number of neurons	It depends on the type of classification that the application needs.
		Activation Function	Only 3 possibilities exist for the output layer. Linear, sigmoid and softmax
Optimizer	The updation of neuron weights from one layer to another is performed using optimizers.		There are 6 types of optimizers available
Loss Function	It defines the model's standard by computing the difference between the actual and predicted values.		In general, there are 9 types of loss functions available. Since the proposed dataset is a multi-classification problem, it uses only 3 types of loss functions.

TABLE III. IMPLEMENTED ACTIVATION FUNCTIONS FOR INPUT AND HIDDEN LAYERS

S.No	Activation Function	Description	Merits
1	Elu	It uses the natural gradients to activate the positive values. It suddenly jumps the mean values [18]	It is low computation, and most values are zero centred
2	Selu	This will always scale the values in such a way that the mean should be zero and the variance should be one	In every iteration, it internally performs the normalization
3	Relu	These output values will never enter into saturation point	The computations are faster
4	Leaky Relu	The model uses simple linear components and adjusts in values in terms of smaller decimal places	It solves the problem of dying Relu
5	Sigmoid	It transforms all the input values between 0 to 1 and produces S- a shape curve	It is easy to differentiate, and prediction values are clear
6	Softplus	It is a derivative of the sigmoid and tanh combination	It always generates smooth curves between -1 to +1
7	Softsign	It transforms the linear values into non-linear values	It is efficient in solving the regression problems
8	Tanh	It is very effective when dealing with negative values and values closer to zero.	When the gradient values are small, then tanh improves the performance.

1) *Activation function*: The activation function is a non-linear change we apply to input before passing it onto the next neuronal layer or finishing it as output [17]. The network can use crucial information and ignore unnecessary data points using activation functions. An activation function takes the values produced by one network level and changes them in a certain way to transfer them to another layer or range of values. The proposed model implements one input layer and three hidden layers. For all these layers, the model passes all the activation functions that suit the dataset as a list of arguments and choose the highest probability as the best parameter. Table III presents the list of activation functions implemented in the proposed model.

### B. Hidden Layers and its Estimators

The proposed model implements all the hidden layers as fully connected layers. The basic principle for any neural network is "the network in which all the hidden layers with the same amount of neurons will have more success rate"[20]. So the proposed model implemented all the hidden layers with the same number of neurons. The hidden layers solve the complex problems by constructing the feature map vector that computes the correlation between each input feature and output class label. In the proposed model, after each hidden layer, it implements an integrated layer that normalizes and drops out the threshold values that are less than the alpha cut. Choosing several hidden layers plays a major role in constructing a neural network. Traditional researchers mentioned that the number of layers should be less than the number of neurons, and also it should be tested from a cross-validation test. Hidden layers are famous for automatic dimensionality reduction, i.e., the number of neurons should be less than the original dimensionality. The below section defines the process of customization.

1) *Customization of neurons count*: The neuron is triggered if the input to an activation function exceeds a threshold; otherwise, it is deactivated. In extremely unusual circumstances, bias will only have one input layer, with the

number of input layer neurons equivalent to the number of features in the data [19]. Bias may rarely have just one input layer, with the number of neurons in the input layer equaling the number of features in the data. At the same time, using the model as a classifier or regressor affects the number of neurons in the output. If the algorithm is a regressor, then the output layer will only include one neuron; however, if the system is a classifier, it may contain one or more neurons, depending on the model's classification label. . As a result, the goal variable affects the count of neurons present in the output layer. The amount of training data, the anomalies, the complexity of data that must be learned from, and the kind of activation functions employed all impact the rate of neurons and layers needed for the hidden layer. Equation 1 presents the computation of the number of optimistic neurons in any layer of the neural network

$$Neurons\_Count = \frac{\text{number of records in dataset}}{\text{bias\_factor} * (\text{input\_neurons} * \text{output\_neurons})} - (1)$$

Fig. 1 denotes integrating hidden layers with drop-out layers to get the normalized values from each layer.

### C. Output Layer and its Estimators

The number of neurons in the output layer equals the number of class labels available in the dataset. Since the proposed model uses a multi-classification dataset, the last layer of a fully connected pattern uses the softmax activation function. It calculates the relative possibilities for all 6 classes and uses an exponential function to normalize the data [21]. This normalization makes the model such that the total 6 class probabilities are summated to 1. But the proposed model uses the softer version of softmax so that all the prediction class label is 1 and other class labels are 0.

### D. Optimizers

An optimizer is a technique or procedure to modify the different parameters that can more efficiently reduce the loss [22]. Different types of optimizers are presented in Fig. 2.

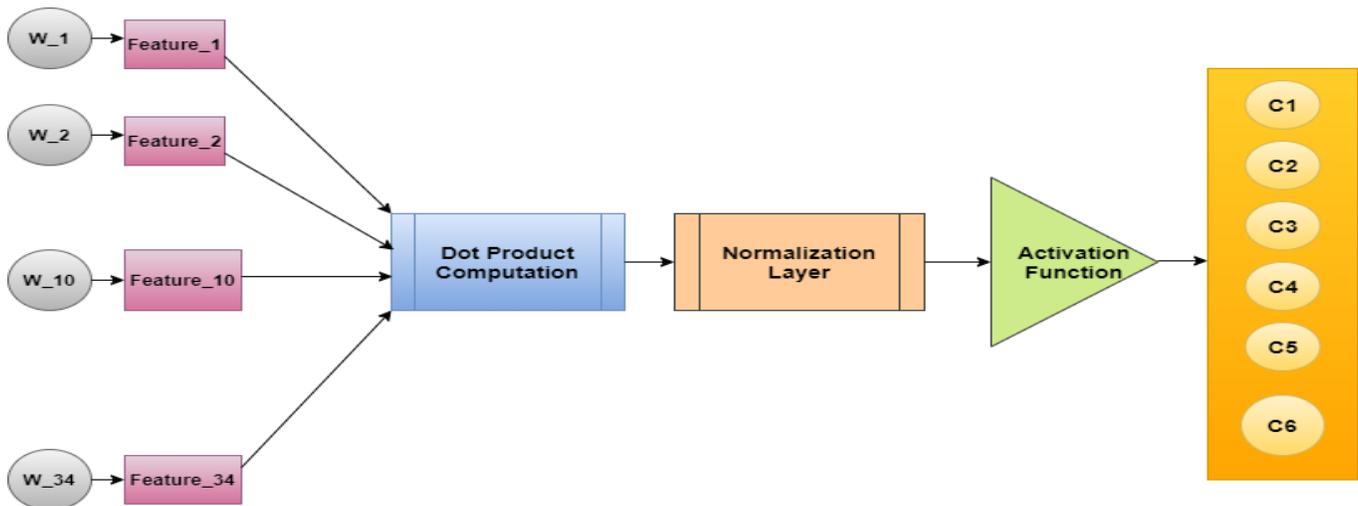


Fig. 1. Integrated architecture of hidden layers.

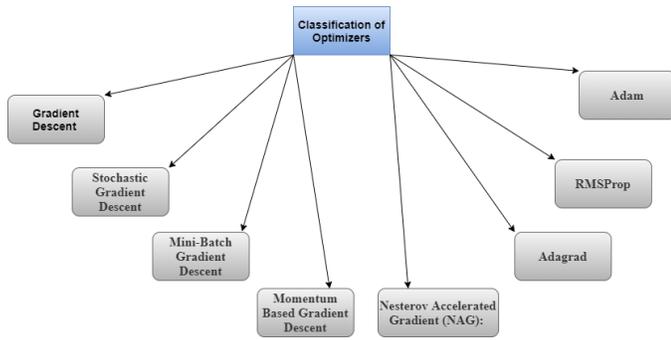


Fig. 2. Classification of optimizers.

1) *Gradient Descent (G.D.)*: Calculus is used by the G.D. optimization method to modify the parameters and consistently find the local minimum. This strategy is also used in neural network backpropagation, where the revised parameters are distributed across the various layers based on when the minimal loss is attained.

2) *Stochastic gradient descent*: Every iteration involves updating the model's parameters. It entails testing the loss function and updating the model after each training sample. Thanks to the stochastic gradient, you can pick the data batches at random. It implies you only have to sample a small part of the dataset.

3) *Mini-batch gradient descent*: Only a bit of the dataset is used in the Mini-Batch Gradient Descent to generate the loss function. As a result, all of the datasets need not be analyzed in memory thanks to batching.

4) *Momentum-based gradient descent*: The gradient descent optimization procedure can ride across flat regions of the search space and overcome the oscillations of noisy gradients by adding momentum. This enables the search to acquire inertia within the search space in a specific direction.

5) *Nesterov Accelerated Gradient (N.A.G.)*: The strategy used in this case was to first update the parameters with the history component before calculating the derivative, which can move the parameters ahead or backwards. This approach, called the look-ahead technique, is more efficient since it can result in fewer oscillations and more time being saved if the curve moves slowly as it approaches the minima.

6) *Adagrad*: The Adagrad optimizer attempts to provide this adaptability by slowing down the training rate according to the modified history of the gradients. The learning rate does not require manual tweaking.

7) *RMSProp*: RPPROP discretely modifies the step size for each weight using the gradient sign. This approach expedites the optimization method by reducing the total of function estimations to find the local minima.

8) *Adam*: Rather than stochastic gradient descent, an alternative optimization approach called Adam can be employed to develop deep learning models. Adam creates an optimization approach that can manage sparse fluctuations in noisy conditions by combining the best elements of both the RMSGrad and AdaProp algorithms.

### E. Loss Functions

The parameter that defines how far the algorithm's current output deviates from the desired output is called the "loss function" [23]. This method is used to judge how well an algorithm matches the data. The parameters that the model learns are established by minimizing a certain loss function, and the loss functions provide a goal against which the model's performance is measured loss functions based on cross-entropy - The distinction between two probabilistic is quantified by cross-entropy. The difference between the probability distribution produced by the activity that produced the data and the distribution that the process model is calculated. The binary cross-entropy is well suited to obtaining one of two outcomes in binary classification scenarios. Multiclass classification uses categorical cross-entropy. In regression circumstances, the model expectation and choices related are real-number values and mean squared error is used. Since the proposed model is a multi-classification problem with six discrete class labels, it implemented "sparse\_categorical\_crossentropy", whose mathematical representation is presented in equation x. since the loss function computes the difference between true and predicted labels let us consider true class labels as  $Y_{true}$  and predicted as  $Y_{predict}$

$$Sparse(Y_{true}, Y_{predict}) = -\frac{1}{n} * \sum_{i=0}^n [y_{true_i} * \log(y_{predict_i}) + (1 - y_{true_i}) * \log(1 - y_{predict_i})] \quad (2)$$

The proposed model considers three estimators as static and assumes all these estimators with the best values. The remaining component estimators are dynamic and are chosen by the optimizer values of the hyper-tuning process. The architecture of the proposed model is presented in Fig. 3.

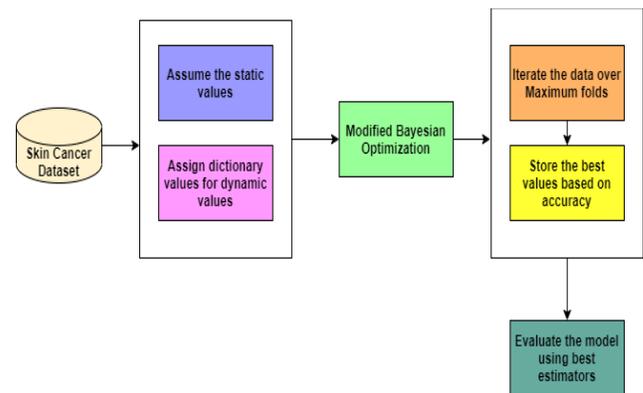


Fig. 3. Workflow of the proposed model.

## V. RESULTS AND DISCUSSION

Table IV denotes the best parameters for every estimator of the component that suits the given dataset for the detection of skin cancer. The proposed model assumes few initial configurations, like it is a designed neural network with three hidden layers, the last dense layer uses softmax function because of the multi-classification, and it implements sparse categorical loss function to evaluate the model.

Fig. 4 denotes the sample screenshot of epochs which have designed the neural network using the best parameters. The training epochs also display the loss and accuracy of training data along with validation loss accuracy. Loss values in almost all iterations are equal to zero for training. So the proposed system is efficient.

Fig. 5 represents the accuracy analysis of existing models with the proposed one to prove state of the art. On X-axis, it denotes the models implemented by the different researchers for identifying skin cancer the Y-axis denotes the percentage accuracies. DCNN has achieved 98% among the existing models. It is the highest accuracy. So when compared to DCNN, the proposed model has achieved +1.65% more.

TABLE IV. BEST ESTIMATORS OF SKIN CANCER DETECTION

S. No.	Estimator Name	Estimator Value
1	Number of neurons in the input layer	47
2	Number of neurons in hidden layer-1	4
3	Number of neurons in hidden layer-2	22
4	Number of neurons in hidden layer-3	24
5	Activation Function for input & hidden layers	Selu
6	Learning Rate	0.46
7	Normalization Rate	0.57
8	Drop Out Rate	0.28
9	Optimizer	Adadelta
10	Batch_size	365
11	Number of epochs	92

```

Epoch 55/92
1/1 [=====] - 0s 51ms/step - loss: 0.0211 - accuracy: 1.0000 - val_loss: 6.6042 - val_accuracy: 0.0278
Epoch 56/92
1/1 [=====] - 0s 51ms/step - loss: 0.0222 - accuracy: 0.9965 - val_loss: 6.5950 - val_accuracy: 0.0278
Epoch 57/92
1/1 [=====] - 0s 52ms/step - loss: 0.0245 - accuracy: 0.9965 - val_loss: 6.6755 - val_accuracy: 0.0278
Epoch 58/92
1/1 [=====] - 0s 53ms/step - loss: 0.0254 - accuracy: 0.9965 - val_loss: 6.7004 - val_accuracy: 0.0278
Epoch 59/92
1/1 [=====] - 0s 49ms/step - loss: 0.0209 - accuracy: 0.9965 - val_loss: 6.7184 - val_accuracy: 0.0278
    
```

Fig. 4. A sample screen shot epochs with best parameters selected.

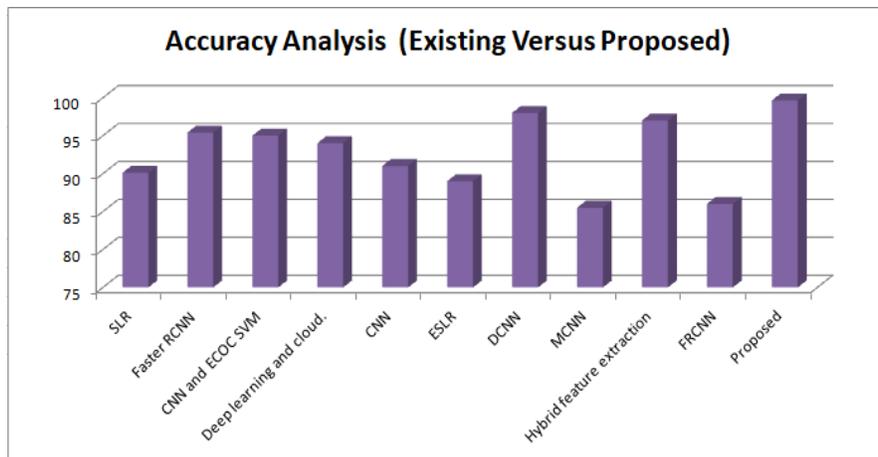


Fig. 5. Evaluation of proposed compared with existing.

iter	target	activa...	batch_...	dropout	dropou...	epochs	layers1	layers2	layers3	learni...	neurons	normal.
2/2	[=====]											
2/2	[=====]											
2/2	[=====]											
2/2	[=====]											
2/2	[=====]											
1	0.3456	5.51	335.3	0.4361	0.3846	43.63	4.58	1.539	11.09	0.2463	40.39	0.9907
2/2	[=====]											
2/2	[=====]											
2/2	[=====]											
2/2	[=====]											
2/2	[=====]											
23	0.874	5.194	364.8	0.2515	0.4846	91.73	3.948	22.15	23.75	0.4653	47.17	0.5771

Fig. 6. 1<sup>st</sup> Best parameters found at 23<sup>rd</sup> iteration.

Fig. 6 represents the iterations performed by the modified Bayesian optimization, in which it sets the iterations to 25 and folds to 2. Out of the 50 iterations, it has achieved its first best values at the 23<sup>rd</sup> iteration and processes the data until it completes all the data processing units. The target is the objective function based on which iteration is best or normal. The higher the target higher the chances of maximization.

Fig. 7 projects the accuracy obtained with the customized ANN by performing 10-fold cross-validation to prove state of

the art. The model designs different layers with different activation functions but with a standard number of neurons. Every layer is designed with popular activation functions, but most of the folds got 100% accuracy, representing overfitting. Finally, the average score for the entire iteration with 10 cross-fold validation is "98.6%", less than the proposed model.

```
Score for fold 10: loss of 0.13818824291229248; accuracy of 97.14285731315613%
-----
Score per fold
-----
> Fold 1 - Loss: 0.16553568840026855 - Accuracy: 97.22222089767456%
-----
> Fold 2 - Loss: 0.016567181795835495 - Accuracy: 100.0%
-----
> Fold 3 - Loss: 0.231260746717453 - Accuracy: 97.22222089767456%
-----
> Fold 4 - Loss: 2.193486398027744e-05 - Accuracy: 100.0%
-----
> Fold 5 - Loss: 0.0038504137191921473 - Accuracy: 100.0%
-----
> Fold 6 - Loss: 0.1421433985233307 - Accuracy: 97.22222089767456%
-----
> Fold 7 - Loss: 0.05570397153496742 - Accuracy: 97.22222089767456%
-----
> Fold 8 - Loss: 0.0027356536593288183 - Accuracy: 100.0%
-----
> Fold 9 - Loss: 0.0007838968886062503 - Accuracy: 100.0%
-----
> Fold 10 - Loss: 0.13818824291229248 - Accuracy: 97.14285731315613%
-----
Average scores for all folds:
> Accuracy: 98.60317409038544 (+- 1.3970062662902984)
> Loss: 0.07567911290152551
-----
```

Fig. 7. Average accuracy score of cross-validation in customized ANN.

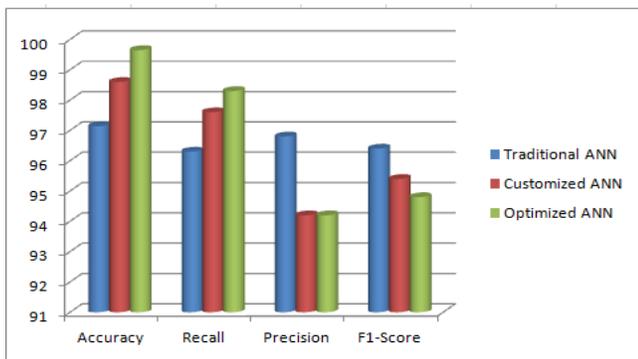


Fig. 8. Metrics evaluation on neural networks algorithms.

Fig. 8 presents the metrics comparison for the three different types of ANN algorithms. Among them optimized ANN has got highest accuracy and recall. X-axis denotes metrics and Y-axis denotes the measurement of metrics.

## VI. CONCLUSION

Identifying Erythemato-Squamous disease using the deep learning approach increases the efficiency of the automation system. The proposed model helps the doctors diagnose the disease with this automation system and predicts the disease at the early stage. Disease identification using the machine learning system has high misclassification and error rates. So few researchers used standard ANNs to train the model with much data and more epochs. But these models are not appropriate because different datasets need different estimators. The proposed model uses optimization techniques to identify the best value for every possible estimator. The model uses a modified version of Bayesian optimization and finds the best values for 11 estimators. The model assumes fixed values like loss functions, number of layers, and number of epochs. Existing optimization techniques need more iterations and memory space. The system to reduce the iterations modifies the Bayesian optimization by comparing the previous iterations and stores only the best values. The model highlights the best values and runs the model till it gets saturation values. Using this model, it has achieved 99.65%. In future work, the model extends the hyper-tuning process by defining a search space, where it can find the parameters depending on the similarity between the attributes and limiting the search space helps the model acquire the learning faster.

## REFERENCES

- [1] Dildar, M., Akram, S., Irfan, M., Khan, H. U., Ramzan, M., Mahmood, A. R., ... &Mahnashi, M. H. (2021). Skin cancer detection: a review using deep learning techniques. *International journal of environmental research and public health*, 18(10), 5479. <https://doi.org/10.3390/ijerph18105479>.
- [2] Nawaz, M., Mehmood, Z., Nazir, T., Naqvi, R. A., Rehman, A., Iqbal, M., & Saba, T. (2022). Skin cancer detection from dermoscopy images using deep learning and fuzzy k-means clustering. *Microscopy Research and Technique*, 85(1), 339-351. <https://doi.org/10.1002/jemt.23908>.
- [3] Dorj, UO., Lee, KK., Choi, JY. et al. The skin cancer classification using deep convolutional neural network. *Multimed Tools Appl* 77, 9909–9924 (2018). <https://doi.org/10.1007/s11042-018-5714-1>.
- [4] Mohammad Ali Kadampur, Sulaiman Al Riyaaee, Skin cancer detection: Applying a deep learning based model driven architecture in the cloud for classifying dermal cell images, *Informatics in Medicine*

Unlocked, Volume 18, 2020, 100282, ISSN 2352-9148, <https://doi.org/10.1016/j.imu.2019.100282>.

- [5] Esteva, A., Kuprel, B., Novoa, R. et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 542, 115–118 (2017). <https://doi.org/10.1038/nature21056>.
- [6] Manne, Ravi, SnigdhaKantheti, and Sneha Kantheti. "Classification of Skin cancer using deep learning, convolutional neural Networks- Opportunities and vulnerabilities-A systematic Review." *International Journal for Modern Trends in Science and Technology*, ISSN (2020): 2455-3778. <https://doi.org/10.1016/j.imu.2019.100282>.
- [7] Hosny, K. M., Kassem, M. A., &Foad, M. M. (2018). Skin Cancer Classification using Deep Learning and Transfer Learning. 2018 9th Cairo International Biomedical Engineering Conference (CIBEC). doi:10.1109/cibec.2018.8641762.
- [8] Daghri, J., Tlig, L., Bouchouicha, M., &Sayadi, M. (2020). Melanoma skin cancer detection using deep learning and classical machine learning techniques: A hybrid approach. 2020 5th International Conference on Advanced Technologies for Signal and Image Processing (ATSIP). doi:10.1109/atsip49331.2020.9231544.
- [9] Vidya, M., & Karki, M. V. (2020). Skin Cancer Detection using Machine Learning Techniques. 2020 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT). doi:10.1109/conecct50063.2020.9198489.
- [10] Jinnai, S.; Yamazaki, N.; Hirano, Y.; Sugawara, Y.; Ohe, Y.; Hamamoto, R. The Development of a Skin Cancer Classification System for Pigmented Skin Lesions Using Deep Learning. *Biomolecules* 2020, 10, 1123. <https://doi.org/10.3390/biom10081123>.
- [11] Mohammad Alnabhan, Ahmad Khader Habboush, Qasem Abu Al-Haija, Arup Kumar Mohanty, SaumendraPattnaik, Binod Kumar Pattanayak, "Hyper-Tuned CNN Using E.V.O. Technique for Efficient Biomedical Image Classification", *Mobile Information Systems*, vol. 2022, Article ID 2123662, 12 pages, 2022. <https://doi.org/10.1155/2022/2123662>.
- [12] Karegowda, A. G., & G., D. (2022). Meta-Heuristic Parameter Optimization for ANN and Real-Time Applications of ANN. In I. Management Association (Ed.), *Research Anthology on Artificial Neural Network Applications* (pp. 166-201). I.G.I. Global. <https://doi.org/10.4018/978-1-6684-2408-7.ch008>.
- [13] H. Alibrahim and S. A. Ludwig, "Hyperparameter Optimization: Comparing Genetic Algorithm against Grid Search and Bayesian Optimization," 2021 IEEE Congress on Evolutionary Computation (C.E.C.), 2021, pp. 1551-1559, doi: 10.1109/CEC45853.2021.9504761.
- [14] Singh, P., Chaudhury, S., &Panigrahi, B. K. (2021). Hybrid MPSO-CNN: Multi-level Particle Swarm optimized hyperparameters of Convolutional Neural Network. In *Swarm and Evolutionary Computation* (Vol. 63, p. 100863). Elsevier B.V. <https://doi.org/10.1016/j.swevo.2021.100863>.
- [15] Pietruszka, M., Borchmann, Ł., &Graliński, F. (2021). Successive Halving Top-k Operator. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 35, Issue 18, pp. 15869–15870). Association for the Advancement of Artificial Intelligence (AAAI). <https://doi.org/10.1609/aaai.v35i18.17931>.
- [16] Sajedi, S., & Liang, X. (2021). Deep generative Bayesian optimization for sensor placement in structural health monitoring. In *Computer-Aided Civil and Infrastructure Engineering* (Vol. 37, Issue 9, pp. 1109–1127). Wiley. <https://doi.org/10.1111/mice.12799>.
- [17] Jagtap, A. D., Shin, Y., Kawaguchi, K., &Karniadakis, G. E. (2022). Deep Kronecker neural networks: A general framework for neural networks with adaptive activation functions. In *Neurocomputing* (Vol. 468, pp. 165–180). Elsevier B.V. <https://doi.org/10.1016/j.neucom.2021.10.036>.
- [18] Apicella, A., Donnarumma, F., Isgrò, F., &Prevete, R. (2021). A survey on modern trainable activation functions. In *Neural Networks* (Vol. 138, pp. 14–32). Elsevier B.V. <https://doi.org/10.1016/j.neunet.2021.01.026>.
- [19] Dey, N., Zhang, Y.-D., Rajinikanth, V., Pugalenth, R., & Raja, N. S. M. (2021). Customized VGG19 Architecture for Pneumonia Detection in Chest X-Rays. In *Pattern Recognition Letters* (Vol. 143, pp. 67–74). Elsevier B.V. <https://doi.org/10.1016/j.patrec.2020.12.010>.
- [20] Muhammad Mazhar Bukhari, Bader Fahad Alkhomees, Saddam Hussain, Abdu Gumaei, Adel Assiri, Syed Sajid Ullah, "An Improved

- Artificial Neural Network Model for Effective Diabetes Prediction", Complexity, vol. 2021, Article ID 5525271, 10 pages, 2021. <https://doi.org/10.1155/2021/5525271>.
- [21] ThangaSelvi, R., Muthulakshmi, I. RETRACTED ARTICLE: An optimal artificial neural network based big data application for heart disease diagnosis and classification model. J Ambient Intell Human Comput 12, 6129–6139 (2021). <https://doi.org/10.1007/s12652-020-02181-x>.
- [22] Abdolrasol M.G.M., Hussain SMS, Ustun TS, Sarker MR, Hannan MA, Mohamed R, Ali JA, Mekhilef S, Milad A. Artificial Neural Networks Based Optimization Techniques: A Review. Electronics. 2021; 10(21):2689. <https://doi.org/10.3390/electronics10212689>.
- [23] Goceri, E. (2021). Diagnosis of skin diseases in the era of deep learning and mobile technology. In Computers in Biology and Medicine (Vol. 134, p. 104458). Elsevier B.V. <https://doi.org/10.1016/j.compbiomed.2021.104458>.
- [24] Attique Khan, M., Akram, T., Sharif, M., Kadry, S., & Nam, Y. (2021). Computer Decision Support System for Skin Cancer Localization and Classification. In Computers, Materials & Continua (Vol. 68, Issue 1, pp. 1041–1064). Computers, Materials and Continua (Tech Science Press). <https://doi.org/10.32604/cmc.2021.016307>.
- [25] W. O'Keefe, B. Ide, M. Al-Khassawneh, O. Abuomar and P. Szczurek, "A CNN Approach for Skin Cancer Classification," 2021 International Conference on Information Technology (ICIT), 2021, pp. 472-475, doi: 10.1109/ICIT52682.2021.9491760.
- [26] Balaji, M.S.P., Saravanan, S., Chandrasekar, M. et al. RETRACTED ARTICLE: Analysis of basic neural network types for automated skin cancer classification using Firefly optimization method. J Ambient Intell Human Comput 12, 7181–7194 (2021). <https://doi.org/10.1007/s12652-020-02394-0>.
- [27] P. S. Silpa et al., "Designing of Augmented Breast Cancer Data using Enhanced Firefly Algorithm," 2022 3rd International Conference on Smart Electronics and Communication (ICOSEC), Trichy, India, 2022, pp. 759-767, doi: 10.1109/ICOSEC54921.2022.9951883.

# Realizing the Quantum Relative Entropy of Two Noisy States using the Hudson-Parthasarathy Equations

Bhaveshkumar B. Prajapati<sup>1</sup> and Nirbhay Kumar Chaubey<sup>2</sup>

Asst. Professor, IT Department, L.D. College of Engineering & Research Scholar, GTU, Gujarat, India<sup>1</sup>  
Senior Member, IEEE

Professor and Dean of Computer Science, Ganpat University, Gujarat, India<sup>2</sup>

**Abstract**—The idea of noisy states can be derived through a quantum relative entropy over a given time period and construct the average value of  $X$  at time based on the system variables. A random Hermitian matrix is used to represent the quantum system observables with BATH states. The Hudson-Parthasarathy (HP equation) context for stochastic processes allows us to simulate quantum relative entropy using quantum Brownian motion. The Sudarshan-Lindblad's density evolution matrix equation was already derivable in generalized form in my previous work. This paper's goal is to illustrate how the HP equation may be used to estimate the density matrix for noise in a perturbed quantum system of a stochastic process. The last stage involves using MATLAB to estimate and simulate a random density matrix and measure the quantum average  $T_r(\rho(t)X)$  at various times. These formulas would be helpful in determining how sensitive the evolving/evolved states are to changes in the Hamiltonian of the noise operators in a sensitivity/robustness study of quantum systems.

**Keywords**—Schrödinger equation; Ito calculus; quantum relative entropy; Hudson-Parthasarathy equation; quantum noise

## I. INTRODUCTION

The origin of Quantum mechanics in 1925, has been expressed to solve many problems and conceived as a generalization of classical mechanics with an added quantum indeterminism [1]. A number of open problems in quantum information theory revolve around whether certain quantities are additive or not. The oldest one was the Holevo capacity method. According to this conjecture, entangled signal states do not improve quantum channel capacity. A second additivity conjecture concerns the minimum entropy of the quantum channel's output [2-4]. As a result, the linear Hilbert space structure is given priority while the probabilistic structure is added almost as an afterthought. It has the unfortunate consequence that in the standard approach to Quantum Mechanics, the dynamical and probabilistic aspects of quantum theory are not quite compatible. There are two distinct modes of wave function evolution: the linear Schrödinger evolution and the probabilistic wave function collapse [5].

A brief but interesting discussion is given on the computation of atomic transition probabilities when the atom interacts with an electromagnetic field. We then calculate using this expression of the atomic state, the average value of an observable on the quantum system as a function of time in terms of the information bearing sequence and use these formulae to derive estimates of the information sequence from a continuous measurement of the observable average [6-9].

After this, we obtain a more accurate description of the measurement and estimation process. When the quantum system is in a pure or mixed state, the measurement of an observable causes the state of the system to collapse to one whose range is contained in the orthogonal eigen-projection of the observable associated with the eigenvalue of  $X$  that has been observed as the outcome. After such a measurement at time  $t_1$ , the system again evolves from the collapsed state under the same Hamiltonian upto time  $t_2 > t_1$  when once again the same observable is measured. Again the state of the system collapses to a state decided by the corresponding eigen-projection and the system evolves from this collapsed state. If  $P$  is the eigen-projection and  $\rho$  is the state just prior to the measurement, then the probability of observing the corresponding outcome is  $T_r(\rho P)$ . In this way, we are able to compute the joint probability of measuring a subset (possibly repeated) of eigenvalues of the observable at a finite set of times  $t_1, \dots, t_N$ , with each time the measurement being made, the system collapsing to a state corresponding to the associated eigen-projection [9-13]. Using quantum measurement models, we examine what kind of measurements can be made on quantum systems, as well as how to determine the probability that a measurement will yield a certain result. In order to find out the effective measurement of quantum states, that technique is very important because there should be a minimum uncertainty of the state with sensitivity to their environment.

One way to describe the output of a single mode, stabilized laser is as a coherent state. An analogy between quantum-mechanical and classical particles oscillating in a harmonic potential is coherent states [5], [7], [11] [24]. Coherent state are eigenstates of the annihilation operator. Coherent states are eigenstates of the annihilation operators' fields. In quantum field theory, creation and annihilation operators' fields are used to correlative the electromagnetic four potential vector field. Thus in a coherent state, one part of the electromagnetic field has defined amplitude and phase. The state  $|n_1, n_2, \dots, n_N\rangle$  correspond to  $n_j$  photons or the  $j^{\text{th}}$  type being present in the bath. Thus  $|n_1, n_2, \dots, n_N\rangle$  is an eigenstates of the number operators  $a_k^\dagger a_k$  with eigenvalue  $n_k$  [8], [10], [19-23].

A bit flip is the only possible error in classical computing while bits are being transferred. Since any rotation or phase shift in Hilbert space represents an error, there are an endless number of distinct faults that could happen for a single qubit in the quantum scenario. During the measurement, a compatible

subspace is projected from the quantum state. When the error is measured, it is brought down to a level that is reasonable for the measurement. In comparison to quantum noise, classical noise has fewer degrees of freedom, making it commutative unitary quantum noise. Classical noise can be created as a particular instance of quantum noise by taking into account the approximate states of both. The novelties of this approach are;

- The time-dependent creation, annihilation, and conservation are introduced in the unique method for quantum stochastic calculus developed by Hudson and Parthasarathy (HP), which satisfies the quantum Ito calculus formula for the product of time differentials of these processes.
- It is demonstrated that the quantum Ito formula of HP naturally evolves into a spectral commutative version of the classical Ito formula for Brownian motion and the Poisson process.
- The Boson Fock space, which is a family of non-commuting operators that specialize to Brownian motion when the state is selected appropriately [12–15], is shown to provide the basis for creating fundamental quantum noise processes in this paper [10–11].
- The linear stochastic model is generalized by HP equations.
- The Schrödinger equation defines a system's unitary evolution when it is coupled to a noisy environment. Because particles can move from the system into the bath and from the bath back into the system, total probability is conserved, which explains why system tensors with BATH exhibit joint unitary evolution [25–29].
- In the HP theory, quantum noise is just a family of non-random operators in Fock Space. When we examine the stochastic linear operator in particular states, randomness appears in all situations. The quantum theory naturally incorporates randomness [14–19]. The classical Ito table is generalized by the quantum Ito tables.

As a result, our main contribution to this study is that infinite-dimensional systems, such as the HP equation, must be truncated in order to achieve a finite-dimensional approximation, which can then be easily reduced utilizing MATLAB through discrimination approaches. We have determined how quickly the respective entropies of the two quantum systems change. Based on their geometric measure of entanglement, some mixed states should allow for the analytical calculation of the rate of change of quantum relative Von Neumann entropy. The principle can be regarded of as a generalization of both the maximum entropy principle and the minimal entropy production principle, both of which are frequently employed in non-equilibrium thermodynamics. This justifies the employment of the principle in the context of optimum learning systems [30–34]. With the use of the symmetric tensor product of a specific Hilbert space, we create the Boson Fock space, which can explain any number of bosons. The Boson Fock space serves as the foundation for

creating fundamental quantum noise processes, such as the noncommuting family of operators that, given the right state selection, specialize to Brownian motion and Poisson processes. A tensor product connects the system Hilbert space to the Boson Fock space, also known as the noise Bath space. Then, we construct the creation, annihilation, and conservation operator fields in the Boson Fock space in accordance with R.L. Hudson and K.R. Parthasarathy's wonderful methodology.

The rest of this paper is written as follows: In Section II, Observable of Quantum Systems using mathematical representations is described. In Section III, the mathematical model of quantum relative entropy for the evolution of two quantum systems is described. In Section IV, the NSER (noise-to-signal energy ratio) to validate the performance criterion is computed. Concluding thoughts are discussed in Section V.

## II. OBSERVABLE OF QUANTUM SYSTEMS

A finite level of a quantum system  $\{A, B\}$  and each system can be described by a finite dimension of Hilbert space  $\{H^A, H^B\}$ . An element of Hilbert space  $H$  is an  $n \times n$  Hermitian matrix with complex entries, called ket vector  $|u\rangle$  and if the same function is linear of the Hilbert space then it is bra vector  $\langle v|$ . The density matrix of a quantum mechanical system is used to compute the mean value of observables. An operator on a Hilbert space with unit trace that is positive semidefinite is called a density operator  $\rho$ . In order for an operator to be considered positive semi definite, it must be Hermitian and have no negative (necessarily real) eigenvalues [5], [22]. Let  $\rho$  is a density matrix of a quantum system and  $X, Y$  two observable on the same Hilbert space. Assume that,  $\text{Tr}(\rho X) = \text{Tr}(\rho Y) = 0$ . Note that  $\text{Tr}(\rho[X, Y])$  is a purely imaginary complex number. A system observable is changed into a system plus noise variable after a finite amount of time by this unitary evolution, which operates on the tensor product of the system and noise Hilbert space. Based on observations made up to time  $t$ , an estimate of this noisy observable at each time  $t$  is required. In order to do this, the measurement process must, however, satisfy the non-demolition property, which requires that the measurement Von Neumann algebra is Abelian and that the measurement at time  $s$  commutes with the state at time  $t$  for time  $t \geq s$ .

Suppose the observable  $X$  evolves in time as  $X(t) = e^{itH} X e^{-itH}$ ,  $t \geq 0$ . Then,  $X(t)$  satisfies the Heisenberg equation of motion for observables:

$$\frac{dX(t)}{dt} = i[H, X(t)].$$

Let  $\rho$  be a density matrix on the same Hilbert space then  $\text{Tr}(\rho(t)X) = \text{Tr}(\rho X(t))$ . For all observables  $X$ , then  $\rho(t) = e^{-itH} \cdot \rho \cdot e^{itH}$  and deduce that,

$$\frac{d\rho(t)}{dt} = i[H, \rho(t)].$$

Therefore, these results can be interpreted in terms of Schrödinger's wave mechanics and Heisenberg's matrix mechanics. Let  $\rho_A$  and  $\rho_B$  be two destiny matrices on  $\mathbb{C}^d$  (both are positive definite with trace one). So, to determine a unitary matrix  $U$  such that  $\|\rho_B - U\rho_A U^*\|$  is a minimum, where  $\|\cdot\|$  denotes Frobenius Norm [23].

Let  $U$  be the optimal unitary matrix. Then for any Hermitian matrix  $H$  must have,

$$\frac{d}{dt} \|\rho_B - Ue^{itH}\rho_A e^{-itH}U^*\|_{t=0}^2 = 0$$

This gives,

$$\text{Tr}((\rho_B - U\rho_A U^*)U[H, \rho_A]U^*) = 0$$

or equivalently,

$$\text{Tr}((U^*\rho_B U - \rho_A)[H, \rho_A]) = 0$$

or

$$\text{Tr}([\rho_A, U^*\rho_B U]H) = 0$$

For all Hermitian matrices  $H$ . It follows that  $U$  must satisfy  $[\rho_A, U^*\rho_B U] = 0$  or  $[U^*\rho_A U, \rho_B] = 0$ . By performing an average over the bath noise variables at each time, we are able to describe how system observables evolve when they are corrupted by bath noise in a way that ensures the system observable always remains a system observable.

### III. MATHEMATICAL MODEL OF QUANTUM RELATIVE ENTROPY

A quantum relative entropy is evolving in between two quantum systems  $\rho_A(t)$  and  $\rho_B(t)$  are density matrices satisfying the Sudarshan- Lindblad equation [33]:

$$\rho'_A(t) = -i[H_A, \rho_A(t)] - \frac{1}{2}\theta_1(\rho_A(t))$$

$$\rho'_B(t) = -i[H_B, \rho_B(t)] - \frac{1}{2}\theta_2(\rho_B(t))$$

$$\text{Where, } \theta_1(X) = \sum_{k=1}^p (L_k^* L_k X + X L_k^* L_k - 2L_k X L_k^*)$$

$$\theta_2(X) = \sum_{k=1}^p (M_k^* M_k X + X M_k^* M_k - 2M_k^* X M_k)$$

Assume  $H_2 - H_1$  and  $M_k - L_k$  upto  $O(\epsilon)$ , then calculate upto  $O(\epsilon^2)$ .

$$\frac{d}{dt} \text{Tr}(\rho_A \log \rho_A) = \text{Tr}\left(\frac{d\rho_A}{dt}\right) + \text{Tr}\left(\rho_A \frac{d}{dt} \log \rho_A\right)$$

$$\text{so by } \rho_A = e^{Z_1}, \rho_B = e^{Z_2},$$

$$\rho'_A = e^{Z_1} \frac{I - e^{-adZ_1}}{adZ_1} (Z'_1)$$

Thus,

$$Z'_1 = \rho_A^{-1} \sum_{r=1}^{\infty} c_r (\text{ad}Z_1)^r (\rho_A^{-1} \rho'_A)$$

$$\text{Tr}(\rho_A \frac{d}{dt} \log \rho_A) = \text{Tr}(\rho_A Z_1)$$

$$= \sum_{r=1}^{\infty} c_r \text{Tr}(\rho_A (\text{ad} \log \rho_A)^r (\rho_A^{-1} \rho'_A))$$

(since  $\text{Tr}(\rho'_A) = 0$ ), so

$$\frac{d}{dt} \text{Tr}(\rho_A \log \rho_A) = \text{Tr}[\rho'_A \log \rho_A] \text{ and,}$$

$$\frac{d}{dt} \text{Tr}(\rho_A \log \rho_B) = \text{Tr}(\rho'_A \log \rho_B) + \text{Tr}(\rho_A Z'_2)$$

$$Z_2 = \log \rho_B, \text{ then } Z'_2 = \frac{\text{ad}Z_2}{1 - e^{-\text{ad}Z_2}} (\rho_A^{-1} \rho'_B)$$

$$\text{Tr}(\rho_A Z'_2) = \text{Tr}\{\rho_A (\text{ad}Z_2 \sum_{m=0}^{\infty} e^{-m \cdot \text{ad}Z_2} (\rho_B^{-1} \rho'_B))\}$$

$$= \sum_{m=0}^{\infty} \text{Tr}(\rho_B^m \rho_A \rho_B^{-m-1} [Z_2, \rho'_B])$$

$$\text{so, } \frac{d}{dt} S(\rho_A, \rho_B) = \frac{d}{dt} \text{Tr}(\rho_A \log \rho_A - \rho_A \log \rho_B)$$

$$= \text{Tr}(\rho'_A \log \rho_A) - \text{Tr}(\rho'_A \log \rho_B) - \text{Tr}(\rho_A Z'_2)$$

$$= \text{Tr}(T_1(\rho_A) \log \rho_A) - \text{Tr}(T_1(\rho_A) \log \rho_B)$$

$$+ \sum_{m=0}^{\infty} \text{Tr}(\rho_B^m \rho_A \rho_B^{-m-1} [\log \rho_B, T_2(\rho_B)])$$

$$\text{Where, } T_k(\rho) = -i[H_k, \rho] - \frac{1}{2}\theta_k(\rho), k = 1, 2 \dots$$

Special case  $\theta_1 = \theta_2 = 0$  (No noise). Then, in this case we find

$$\frac{d}{dt} S(\rho_A, \rho_B) = i \text{Tr}\{(H_A - H_B)[\rho_A, \log \rho_B]\}$$

When it comes to general terms  $\rho_B = \sum_{\alpha=1}^p p_{\alpha} |e_{\alpha}\rangle \langle e_{\alpha}|$  is the spectral representation of  $\rho_B$  with  $p_{\alpha} > 0, \forall \alpha$ . Then let  $X = [\rho_A, \log \rho_B]$ , we get

$$\rho_B^N [\rho_A, \log \rho_B] \rho_B^N =$$

$$\sum_{\alpha, \beta=1}^p (p_{\alpha}/p_{\beta})^N |e_{\alpha}\rangle \langle e_{\beta}| \langle e_{\alpha}| X |e_{\beta}\rangle$$

If we assume that  $p_{\alpha} > p_{\beta} \Rightarrow \langle e_{\alpha}| X |e_{\beta}\rangle = 0$  then,

$$\lim_{N \rightarrow \infty} \rho_B^N [\rho_A, \log \rho_B] \rho_B^N =$$

$$= \sum_{\alpha, \beta: p_{\alpha} = p_{\beta}} |e_{\beta}\rangle \langle e_{\alpha}| X |e_{\beta}\rangle = X$$

and we then get

$$\frac{d}{dt} S(\rho_A, \rho_B)$$

$$= i \text{Tr}\{(H_A - H_B)X\} + i \text{Tr}\{H_B X\}$$

We note that

$$X = [\rho_A(t), \log \rho_B(t)] \equiv X(t)$$

$$= [U_1(t)\rho_A(0)U_1^*(t), U_2(t)\log \rho_B(0)U_2^*(t)]$$

$$= U_1(t)\rho_A(0)U_1^*(t)U_2(t)\log \rho_B(0)U_2^*(t)$$

$$- U_2(t)\log \rho_B(0)U_2^*(t)U_1(t)\rho_A(0)U_1^*(t)$$

Where,  $U_1(t) = \exp(-itH_A), U_2(t) = \exp(-itH_B)$

Now assume,  $U_1^*(t)U_2(t) \xrightarrow{t \rightarrow \infty} \Omega$ , (scattering matrix)

$$\begin{aligned} \text{Then, } \lim_{t \rightarrow \infty} T_r((H_A - H_B)X(t)) \\ = \lim_{t \rightarrow \infty} T_r\{U_2^*(t)U_1(t)(H_A - H_B)U_1(t)\rho_A(0)\Omega \log \rho_B(0)\} \end{aligned}$$

Now,

$$\frac{d}{dt} U_2^*(t)U_1(t) = -iU_2^*(t)(H_A - H_B)U_1(t)$$

$$\text{Write, } \Omega = \Omega(\infty) = \lim_{t \rightarrow \infty} U_1^*(t)U_2(t);$$

$$\Omega(t) = U_1^*(t)U_2(t).$$

$$\text{Then, } U_2^*(t)(H_A - H_B)U_1(t) = -i\Omega'(t),$$

$$\text{and we get, } \lim_{t \rightarrow \infty} T_r((H_A - H_B)X(t))$$

$$= i T_r\{\Omega'(\infty)\rho_A(0)\Omega(\infty)\log \rho_B(0)\}.$$

If  $\Omega'(\infty) = 0$  then this vanishes and we get,

$$\lim_{t \rightarrow \infty} \frac{d}{dt} S(\rho_A(t), \rho_B(t)) = i T_r\{H_B X(\infty)\}$$

We've seen that

$$\frac{d}{dt} S(\rho_A, \rho_B) = i T_r\{H_A[\rho_A, \log \rho_B]$$

$$+ i \lim_{N \rightarrow \infty} T_r\{H_B \rho_B^N[\rho_A, \log \rho_B] \rho_B^{-N}\}$$

$$- i T_r\{H_B[\rho_A, \log \rho_B]\}$$

$$\text{Let, } \Omega(t) = U_1^*(t)U_2(t).$$

Then,

$$\begin{aligned} & [\rho_A, \log \rho_B] \\ &= U_1(t)\rho_A(0)U_1^*(t)\log \rho_B(0)U_2^*(t) \\ &= U_1(t)\rho_A(0)\Omega(t)\log \rho_B(0) \\ & \quad U_2^*(t) \end{aligned}$$

So,

$$i T_r\{(H_A - H_B)[\rho_A(t), \log \rho_B(t)]\}$$

$$= i T_r\{U_2^*(t)(H_A - H_B)U_1(t)\rho_A(0)\Omega(t)\log \rho_B(0)\}$$

$$= -T_r\{\Omega'(t)\rho_A(0)\Omega(t)\log \rho_B(0)\}$$

and,

$$T_r\{H_B \rho_B(t)^N[\rho_A(t), \log \rho_B(t)]\rho_B(t)^{-N}\}$$

$$= T_r\{H_B U_2(t)\rho_B(0)^N U_2^*(t)U_1(t)\rho_A(0)$$

$$U_1^*(t)U_2(t)\log \rho_B(0)\rho_B(0)^{-N} U_2^*(t)\}$$

$$- T_r\{H_B U_2(t)\rho_B(0)^N \log(\rho_B(0))U_2^*(t)$$

$$U_1(t)\rho_A(0)U_1(t)U_2(t)\rho_B(0)^{-N} U_2^*(t)\}$$

$$= T_r\{H_B \rho_B(0)^N \Omega(t)\log \rho_B(0)\rho_B^{-N}$$

$$- T_r\{H_B \rho_B^N \log \rho_B(0)\Omega^*(t)\rho_A(0)\Omega(t)\rho_B^{-N}\}$$

Also note that

$$T_r\{U_2^*(t)(H_A - H_B)$$

$$U_1(t)\rho_A(0)\Omega(t)\log \rho_B(0)\}$$

$$\begin{aligned} &= T_r\{U_2^*(t)U_1(t)H_A \\ & - H_B U_2^*(t)U_1(t)\rho_A(0)\Omega(t)\log \rho_B(0)\} \\ &= T_r\{(\Omega^*(t)H_A \\ & - H_B \Omega^*(t))\rho_A(0)\Omega(t)\log \rho_B(0)\} \end{aligned}$$

So,

$$\frac{d}{dt} S(\rho_A(t), \rho_B(t)) =$$

$$\lim_{N \rightarrow \infty} i \lim_{N \rightarrow \infty} T_r\{H_B \rho_B(0)^N [\Omega^*(t)\rho_A(0)\Omega(t)\log \rho_B(0)] \rho_B^{-N}(0)\}$$

$$+ i T_r\{(\Omega^*(t)H_A - H_B \Omega^*(t))\rho_A(0)\Omega(t)\log \rho_B(0)\}$$

Let,  $\rho_B(0) = \sum_{\alpha=0}^r p_\alpha(\alpha)p_\alpha$  be the spectral distinct positions of  $\rho_B(0)$ . Thus  $\{\rho_A(0), \dots, \rho_r(0)\}$  are distinct,  $\sum_{\alpha} p_\alpha(0)T_r(p_\alpha) = 1$ ,  $\sum_{\alpha} p_\alpha = I$ ,  $P_\alpha P_\beta = P_\alpha \delta_{\alpha\beta}$ ,  $P_\alpha^* = P_\alpha$ .

In this paper, the rate of change entropy of the two quantum systems is solved and the parameters of the Hamiltonians of the noise operators are determined, which will yield the exact value of the relative entropy of entanglement.

Let us consider,

$$\rho'_A(t) = -i[H_A, \rho_A(t)] \text{ and}$$

$$\rho'_B(t) = -i[H_A, \rho_B(t)] - \frac{1}{2}\theta(\rho_B(t))$$

Where,

$$\theta(X) = L * LX + XL * L - 2LXL *$$

Then,

$$\frac{d}{dt} T_r(\rho_A \log \rho_A) = 0$$

$$\frac{d}{dt} T_r(\rho_A \log \rho_B)$$

$$= T_r(\rho'_A \log \rho_B) + T_r(\rho_A Z'_2)$$

$$Z'_2 = \frac{adZ_2}{1 - e^{-adZ_2}} (\rho_B^{-1} \rho'_B)$$

$$= adZ_2 \sum_{m=0}^{\infty} e^{-m.adZ_2} (\rho_A^{-1} \rho'_B)$$

$$= adZ_2 \left( \sum_{m=0}^{\infty} \rho_B^{-m-1} \rho'_B \rho_B^m \right)$$

So,

$$T_r(\rho_A, Z'_2) = \sum_{m \geq 0} T_r(\rho_A \rho_B^{-m-1} [Z_2, \rho'_B] \rho_B^m)$$

$$= -i \sum_{m \geq 0} T_r(\rho_A \rho_B^{-m-1} [Z_2, [H_B, \rho_B]] \rho_B^m)$$

$$- \frac{1}{2} \sum_{m \geq 0} T_r(\rho_A \rho_B^{-m-1} [Z_2, \theta(\rho_B)] \rho_B^m)$$

$$\begin{aligned} & \frac{d}{dt} T_r(\rho_A \log \rho_B) = \\ & T_r(\rho'_A \log \rho_B) - \iota \sum_{m \geq 0} T_r(\rho_A \rho_B^{-m-1} [Z_2, [H_B, \rho_B]] \rho_B^m) \\ & - \frac{1}{2} \sum_{m \geq 0} T_r(\rho_A \rho_B^{-m-1} [Z_2, \theta(\rho_B)] \rho_B^m) \end{aligned}$$

In conclusion, the change in relative entropy of the two quantum systems is given below;

$$\begin{aligned} & \frac{d}{dt} S(\rho_A, \rho_B) = \\ & \frac{d}{dt} T_r(\rho_A \log \rho_A - \rho_A \log \rho_B) \\ = & T_r(\rho'_A \log \rho_A) - T_r(\rho'_A \log \rho_B) - T_r(\rho_A Z'_2) \\ & \frac{d}{dt} S(\rho_A, \rho_B) = -T_r(\rho'_A \log \rho_B) \\ & + \iota \sum_{m \geq 0} T_r(\rho_A \rho_B^{-m-1} [Z_2, [H_B, \rho_B]] \rho_B^m) \\ & + \frac{1}{2} \sum_{m \geq 0} T_r(\rho_B \rho_B^{-m-1} [Z_2, \theta(\rho_B)] \rho_B^m) \end{aligned}$$

Where,

$$\theta(\rho_B) = L * L \rho_B + \rho_B L * L - 2L \rho_B L *$$

In order to calculate in time, the initial signal of observable X, the noisy Schrödinger equation is simulated by using a large set of exponential vectors in the Boson-Fock space for noisy baths, and an orthogonal basis for the signal Hilbert space.

#### IV. RESULTS AND DISCUSSIONS

The observable with the ideal theta value is shown in this section, along with a graphic of the noise to signal energy ratio.

##### A. The Expectation Value of an Observable of Two Quantum System

This paper is used to determine the average value X, which is expressed by  $X = T_r(\rho X)$  of the state of a quantum-mechanical system as described by its density operator  $\rho$  with  $T_r(\rho) = 1$ . The definition of the observable X, which is a  $2 \times 2$  random Hermitian matrices, as well as the values of  $H_0$  and  $V$ , constitute the first step in the analysis. It computes this matrix's eigenvalues and accompanying eigenvectors.

By comparing the expected value of an observable in a mixed state with the expected value of the observable in several mutually orthogonal pure states, one can obtain the expected value of the observable in the mixed state [12-15]. Our results are explaining the quantum average measure value to extra the information with effecting of the AWGN and stochastic noise and we will evaluate the performance of our algorithm in the presence of BATH states, that is, compute the noise to signal ratio of the given estimate of  $\delta\theta$ , that is,  $E(|\delta(\theta) - \delta'(\theta)|^2)$ . Using functional analysis and strict mathematics, it is possible to generate the quantum noise. In particular, we demonstrate how some important stochastic processes from classical probability theory, such as the Brownian motion and Poisson processes, can be viewed as special cases of quantum stochastic processes, which are a family of non-commuting observables in a particular type of Hilbert space called the

Boson Fock space when observed in particular states. The randomly generated two Hamiltonian of the given system and find the Eigenvalues through MATLAB.

---

##### Algorithm 1: An algorithm of the expectation value of an observable of two quantum system

---

```
Data: observable X ≥ 0
Result: X = Tr(ρX)
Density operator = ρ;
Tr(ρ) = 1;
H0 ← Hermitian Operator;
V ← Hermitian Operator;
Taking the initial state ψ
ψ0 = rand(3,1) + i * rand(3,1)
For making norm = 1 of ψ
φ ← choose
ψ0 =  $\frac{\psi_0}{\text{norm}(\psi_0)}$ 
sum ψ = zeros(2, 2, N)
A1 ← choose
A2 ← choose
P1 ← choose
P2 ← choose
I = 0 ← choose
Q = (A1 ⊕ P1) + (P1 ⊕ A1) + (A2 ⊕ P2) + (P2 ⊕ A2) +
(A3 ⊕ P3) + (P3 ⊕ A3)
while N ≠ 0 do
    if N is integer value of qubit then
        X ← Tr(X* × X) ← minimum;
        θ = (0.5 * Tr(A * A') * real(f1 * f'1) *
real(f1 * f'1) * Tr(A2))-1 * real(f1 * Tr(A * Q))
        is Minimum
    end
    otherwise: δθ =  $\frac{\text{real}(T_r(\delta P * Q))}{T_r(Q * Q')}$  ← estimating
    is Minimum
end
```

---

Where,

$$\delta P = \delta \theta * Q$$

And,

$$\begin{aligned} H_A &= \text{rand}(2,2) + j * \text{rand}(2,2); \\ H_A &= [ 0.8178 \quad 0.4275 - 0.0756i; \\ & \quad 0.4275 + 0.0756i \quad 0.0225] \end{aligned}$$

The Hermitian matrix equation of the Hamiltonian of the first system is given below;

$$\begin{aligned} H_A &= (H_A + H_A') / 2 \\ H_B &= \text{rand}(2,2) + j * \text{rand}(2,2); \\ H_B &= [ 0.4229 \quad 0.3464 - 0.0307i; \\ & \quad 0.3464 + 0.0307i \quad 0.4709] \end{aligned}$$

$$\delta X = \begin{array}{|c|c|c|c|} \hline \begin{array}{c} -0.070838082586555 \\ - 0.000000000000000i \end{array} & \begin{array}{c} -0.015815573941431 \\ + \\ 0.007750834345286i \end{array} & \begin{array}{c} -0.015815573941431 \\ + \\ 0.007750834345286i \end{array} & \begin{array}{c} 0.071195946682088 - \\ 0.010664577493442i \end{array} \\ \hline \begin{array}{c} -0.015815573941431 \\ - 0.007750834345286i \end{array} & \begin{array}{c} 0.070838082586553 + \\ 0.000000000000000i \end{array} & \begin{array}{c} 0.070838082586553 + \\ 0.000000000000000i \end{array} & \begin{array}{c} 0.015815573941431 - \\ 0.007750834345286i \end{array} \\ \hline \begin{array}{c} -0.015815573941431 \\ - 0.007750834345286i \end{array} & \begin{array}{c} 0.070838082586553 + \\ 0.000000000000000i \end{array} & \begin{array}{c} 0.070838082586553 + \\ 0.000000000000000i \end{array} & \begin{array}{c} 0.015815573941431 - \\ 0.007750834345286i \end{array} \\ \hline \begin{array}{c} 0.071195946682088 \\ + \\ 0.010664577493442i \end{array} & \begin{array}{c} 0.015815573941431 + \\ 0.007750834345286i \end{array} & \begin{array}{c} 0.015815573941431 + \\ 0.007750834345286i \end{array} & \begin{array}{c} -0.070838082586553 \\ + \\ 0.000000000000000i \end{array} \\ \hline \end{array}$$

The Hermitian Matrix equation of the Hamiltonian of the second system is given below;

$$H_B = (H_B + H_B^*) / 2;$$

First, initialize the states and choose the value of  $A_1, A_2, P_1$  and  $P_2$  for estimating the theta. So, the estimated value of  $\theta$  is 0.33 for  $2 \times 2$  matrices, additionally, if we increase the size of the qubit, the estimated values of  $\theta$  is 0.28 for  $3 \times 3$ . Then second, we are designed  $\delta X$  for  $2 \times 2$  matrices, consider  $\delta\theta = 0.33$  with a random noise generating from random AWGN.

Since the collapse postulate is taken into account, continuous measurement is not covered in this section. A single measurement is known to cause the system's state to collapse to the eigenstate of the measured observable, which corresponds to the observed result. Since the metric above only displayed the minimal value of the observable, it is clear that, for each time index of T, the value of our error energy function approaches zero or its smallest value.

**B. NSER (Noise- to-Signal Energy Ratio) of Entropy**

We can use the NSER as a performance index to calculate the impact of this noise on the error energy. This chart, which is displayed below, shows how noise to signal energy ratios can remain low over an extended length of time, which verifies a successful performance. Using the Frobenius norm, we must minimize the error function or cost function.

The error function has to be determined the optimum operators  $L_1, L_2, S$ , so that  $\int_0^T \|\rho_s(t, u) - \rho_d(t)\|^2 dt$  is a minimum. Assuming the function  $u(t), 0 \leq t \leq T$  and H given.

$$NSER = \frac{\|\rho_s(t, u) - \rho_d(t)\|^2}{\rho_d(t)^2} = \frac{\zeta_{min}}{\rho_d(t)^2}$$

**Algorithm 2:** NSER (Noise – to – Signal Energy Ratio) of entropy

**Data:** Using the Frobenius norm, we must minimize the error function or cost function.

**Result:** NSER should be remains less than unity. initialization;

```

while While condition do
    instructions;
    if condition then
        instructions1:  $\delta\theta$  is minimum;
        instructions1:  $\theta$  is minimum;
        then
            X ← minimum
        then
            instructions3:  $\rho$  ← minimum
    then NSER =  $s(t, u) - d(t)^2 / (d(t)) = min / (d(t))$ 
    else
        instructions4: NSER is minimum;
    Outcome: Rate of the change of Entropy is minimum
    Final:
         $\frac{dS(\rho_A - \rho_B)}{dt} = \frac{dT_{r(\rho_A \log(\rho_A) - \rho_B \log(\rho_B))}}{dt}$ 
    end
end

```

Where,  $\zeta_{min}$  is the error energy function value and  $\rho_d(t)$  is the desired density function. We show through simulation that the NSER of entropy stabilizes to a small value, supporting the information inequality that states conditionally reducing entropy decreases information. A better design, one with a lower SER, might theoretically be obtained by first averaging over the noise distribution and then minimizing with respect to the nonrandom functions (see Fig. 1).

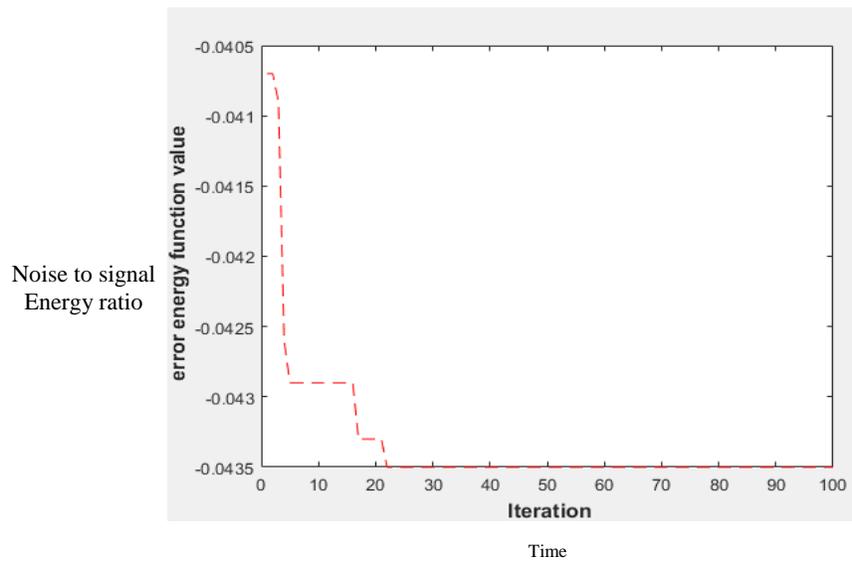


Fig. 1. Noise-to-signal energy ratio plot for quantum system

Since reducing noise effects and highlighting the signal process is the whole purpose of collecting measurements, the NSER should gradually decrease with time.

$$\frac{d}{dt}S(\rho_A, \rho_B) = \frac{d}{dt}T_r(\rho_A \log \rho_A - \rho_A \log \rho_B)$$

We compute NSER, and it remains less than unity. We also showed how to use stochastic differential equations to calculate the relative entropy of two quantum systems plus bath density, i.e. stochastic system density. Through simulations, we are justifying that the rate of change of relative entropy stabilizes to a value less than  $\theta(\rho_B)$ , which realistically justifies the information inequality stating that conditioning reduces entropy. The idea will be helpful to research communities in applied mathematics, physics, and quantum information theory who seek to investigate the variety of applications of classical and quantum stochastics to issues of physics and engineering, to sum up the conclusion.

## V. CONCLUSIONS

We have determined the quantum relative entropy rate between two mixed states using noisy Schrödinger equations with varied Hamiltonians and Lindblad operators. For our calculations, we applied the conventional formula for calculating the exponential map of matrices. We may compute the rate of relative entropy as the asymptotic limit  $t \rightarrow \infty$  based on the scattering matrices connected to the pair of Hamiltonians that generate the two states. It is important to look into the circumstances in which the asymptotic relative entropy rate for the Hamiltonian and Lindblad operators continues to be below the specified threshold. The asymptotic limit in this scenario would ensure that there is a short gap between the two states. Conditioning is known in classical information theory to decrease entropy, specifically  $H(X/Y) \leq H(X)$ . As a result, we anticipate that in the quantum setting, the entropy of the filtered state will be reduced using HP equations based on detecting the noise process. In further work, we will also extend this formalism to Belavkin's quantum filtering

theory based on the Hudson-Parthasarathy quantum stochastic by demonstrating that when this equation for a particle travelling in a potential with damping and noise is characterized in terms of the Wigner distribution function, then it is exactly the same as the Kushner-Kallianpur stochastic filter but with quantum correction terms stated as a power series in Planck's constant.

## REFERENCES

- [1] Shapiro J.H., Zhang Z., Wong F.N.C.: Secure communication via quantum illumination. *Quantum Information Processing*, 13(10), (2014)
- [2] Dirac P.A.M.: *The principles of quantum mechanics*, Oxford university press, 4th edition, New York, (1958)
- [3] Schrödinger E.: An Undulatory Theory of the Mechanics of Atoms and Molecules. *Phys. Rev.*, 28,(1926)
- [4] Griffiths D.J.: *Introduction to quantum mechanics*, Pearson Hall, 2nd edition, USA, (2005).
- [5] Christopher A. F., Carlton M. C.: Mathematical techniques for quantum communication theory. *Open Systems and Information Dynamics*, 3(3), 345-356(1995)
- [6] Garg N., Parthasarathy H., Upadhyay D.K.: "Real-time simulation of H-P noisy Schrödinger equation and Belavkin filter", *Quantum Information Processing*, 10.1007/s11128-017-1572-4, (2017).
- [7] Von Neumann J.: *Mathematical Foundations of Quantum Mechanics*, Princeton Univ. Press, Princeton, NJ, (1955).
- [8] Helstrom C.W.: *Detection Theory and Quantum Mechanics*. *Information and Control*, 10, 254-291(1967)
- [9] Helstrom C.W.: *Detection Theory and Quantum Mechanics (II)*. *Information and Control*, 13, 156-171(1968)
- [10] Peyton B., Sard E., Lange R., Arams F.R.: *On Optimal Quantum Receivers for Digital Signal Detection*. *Proceedings of the IEEE*, (1970)
- [11] Li K., Zuo Y., Zhu B.: Suppressing the Errors Due to Mode Mismatch for M-ary PSK Quantum Receivers Using Photon-Number-Resolving Detector. *IEEE Photon. Technol. Lett.*, 25(22),(2013)
- [12] Shi J., Shi R., Guo Y., Peng X., Lee M. H.: Probabilistic quantum relay communication in the noisy channel with analogous space-time code. *Quantum Information Processing*, 12, 1859-1870(2013)
- [13] Kato K., Hirota O.: Square-Root Measurement for Quantum Symmetric Mixed State Signals. *IEEE Transaction on Information Theory*, 49(12), (2003)

- [14] Atmanspacher H., Kurths J., Scheingraber H., Wackerbauer R., Witt A.: Complexity and meaning in nonlinear dynamical systems, *Open Systems and Information Dynamics*, 1(2), 269-289(1992)
- [15] Holevo A.S.: Statistical Decision Theory for Quantum Systems. *Journal Of Multivariate Analysis*, 3, 337-394(1973)
- [16] Yuen H.P., Kennedy R.S., Lax M.: Optimum Testing of Multiple Hypotheses in Quantum Detection Theory. *IEEE Transaction on Information Theory*, 21(2), (1975)
- [17] Helstrom C.W.: Bayes-Cost Reduction Algorithm in Quantum Hypothesis Testing. *IEEE Transaction on Information Theory* 28(2), 359-366 (1982)
- [18] Kennedy R.S.: A near-optimum receiver for the binary coherent state quantum channel. *Research Lab. Electron MIT Cambridge Tech. Rep.*, 108, (1973)
- [19] Dolinar S.J.: An optimum receiver for the binary coherent state quantum channel. *Research Lab. Electron MIT Cambridge Tech. Rep.*, 111, (1973)
- [20] Sasaki M., Hirota O.: Optimum decision scheme with a unitary control process for binary quantum-state signals. *Phys. Rev. A*, 54, (1996)
- [21] Wang D., Huang A.J., Sun W.Y., Shi J.D., Ye L.: "Practicle single-photon-assisted remote state preparation with non-maximally entanglement", *Quantum Information Processing*, (15), 3367-3381, (2016).
- [22] Wang, D., Hu, Y. D., Wang, Z. Q., Ye, L.: "Efficient and faithful remote preparation of arbitrary three-and four-particle W-class entangled states". *Quantum Information Processing*, 14(6), 2135-2151, (2015).
- [23] Lo H.K.: "Classical-communication cost in distributed quantum-information processing: a generalization of quantum-communication complexity" *Physical Review A*, (62), (2000).
- [24] Wang D., Hoen R.D., Ye L., Kais S.: "Generalized Remote Preparation of Arbitrary  $m$ -qubit Entagled States via Genuine Entanglements", *Entropy* (17), 1755-1774, (2015)
- [25] Hayashi M., Universal Coding for Classical-Quantum Channel, *Communication in Mathematical Physics* 289, 1087-1098, (2009)
- [26] Nielsen, M.A., Chuang, I.L.: *Quantum Computation and Quantum Information*, pp. 171–286. Cambridge University Press, Cambridge (2001).
- [27] Parthasarathy K.R.: "Coding theorems of classical and quantum Information theory", Hindustan Book Agency; 2013.
- [28] Dowker F., Tabatabai Y.G.: Dynamical wavefunction collapse models in quantum measure theory. *J. Phy. A Mathematical and Theoretical*, 41(20), (2008)
- [29] Mensky M., Audretsch J.: Continuous QND measurements: no quantum noise. *Journal Applied Physics B*, 64(2), 129-136, (1997)
- [30] Srikanth R.: A Computational Model for Quantum Measurement. *Quantum Information Processing*, 2(3), 153-199, (2003)
- [31] Gisin N., Popescu S., Scarani V., Wolf S. and Wullschleger J.: Oblivious transfer and quantum channels as communication resources. *Natural Computing*, 12(1), 13-17, (2013).
- [32] Karlsson A., Björk G.: Quantum correlations in dual quantum measurements. *Journal Applied Physics B*, 64(2), 235-241, (1997)
- [33] Prajapati, Bhaveshkumar B. and Chaubey, N. K., "Realization of Relative Entropy Evolution in the Sudarshan-Lindblad for two Quantum Systems", *JOURNAL OF ALGEBRAIC STATISTICS*, Volume 13, No. 1, 2022, p. 490-497.
- [34] Prajapati, Bhaveshkumar B. and Chaubey, N. K., "Quantum Key Distribution: The Evolution" *Quantum Crypstography and the Future of Cyber Security (IGL Global)*, 2020 |P.: 15.

# Research on Automatic Detection Algorithm for Pedestrians on the Road Based on Image Processing Method

Qing Zhang

Zhangjiajie College, Jishou University, Zhangjiajie, Hunan 427000, China

**Abstract**—Accurate detection of pedestrian targets can effectively improve the performance level of intelligent transportation and surveillance projects. In order to effectively enhance the accuracy of detecting pedestrian targets on the road, this paper first introduced the traditional pedestrian target detection algorithm, proposed the faster recurrent convolutional neural network (RCNN) algorithm to detect pedestrian targets, and improved it to make good use of the convolutional features at different scales. Finally, support vector machine (SVM), traditional Faster RCNN, and optimized Faster RCNN algorithms were compared by simulation experiments. The results showed that the optimized Faster RCNN algorithm had higher detection accuracy and recall rate, obtained a more accurate target localization frame, and detected faster than SVM and traditional Faster RCNN algorithms; the traditional Faster RCNN algorithm had higher detection accuracy and target frame localization accuracy than the SVM algorithm.

**Keywords**—Pedestrian detection; recurrent convolutional neural network; scale-invariant feature transform; support vector machine; characteristic scale; Difference of Gaussians operator

## NOMENCLATURE

$D(x, y, \sigma)$ : The DoG operator.

$DR$ : The target frame predicted by the algorithm.

$G(x, y, \sigma)$ : The Gaussian filter function.

$G(x, y, k\sigma)$ : The Gaussian filter function.

$GT$ : The actual target frame in the image.

$IoU$ : The degree of target frame overlap.

$I(x, y)$ : The original image.

$P$ : The precision.

$R$ : The recall rate.

## I. INTRODUCTION

Economic development continues to improve people's living standards, and the pressure on traffic management increases as more and more vehicles are used in travel [1]. The progress of computer technology has promoted the emergence of intelligent transportation, and the detection of pedestrians is an important component of intelligent transportation [2]. Accurate detection of pedestrians can effectively improve the level of intelligent driving, intelligent monitoring, and other

technologies. For example, in intelligent driving, more accurate pedestrian detection can assist drivers to make safe avoidance of pedestrians and reduce the occurrence of traffic accidents; in intelligent monitoring, computers replace humans to make recognition of pedestrians in monitoring videos, track pedestrians, and judge the behavior of pedestrians, thus improving the security level. Manual identification is relatively accurate and is also more intuitive when identifying pedestrians in video images, but human energy is limited and cannot maintain focused attention for a long time, so replacing humans with machines to automatically detect pedestrians is the current trend. Although the detection of pedestrians in images by image processing techniques is not intuitive, it is relatively more comprehensive in measuring targets with smaller scales in images. The traditional pedestrian target detection algorithm uses a feature extraction algorithm to extract image features before classification by a classification algorithm. In the traditional pedestrian target detection algorithm, feature extraction and recognition and detection of images can be considered relatively independent, and the features extracted by the feature extraction algorithm are often statistical local features, which are difficult to reflect image features comprehensively. As deep learning algorithms and computer performance improve, convolutional neural networks (CNNs) have been applied to pedestrian detection. Compared with the traditional detection method, CNNs combine image feature extraction and recognition together and integrated the local features extracted using convolutional kernels into global features, thus making the detection of pedestrian targets more accurately.

Some relevant literature is reviewed below. Xu et al. [3] reconstructed a target detection model called YOLOv3, proposed YOLOv3-promote, and introduced an attention mechanism. They found that the inference speed of the method was faster than the original model and the parameter volume was reduced to one-tenth. Xia et al. [4] put forward a pedestrian detection algorithm based on multi-scale feature extraction and attention feature fusion and found that the algorithm had good detection performance. Liu [5] proposed a deep residual network-based adaptive scale pedestrian detection algorithm and found that the algorithm was applicable to pedestrians of different scales. Yang et al. [6] designed a pedestrian target detection algorithm based on a single shot multibox detector. Subsequent simulation experimental results on VOC2007 and data\_sub showed that the maximum value of mAP was 77% and the maximum

accuracy was 96.31%. Zhang et al. [7] designed a pedestrian target detection algorithm based on the histogram of oriented gradient and support vector machine (SVM). They found that the algorithm greatly reduced the computational effort when feature extraction was performed only on candidate regions, thus improving the detection efficiency. Pei et al. [8] designed a multispectral pedestrian target detection algorithm combining visual optical images and infrared images based on deep CNNs and performed simulation tests on the public multispectral benchmark dataset. They found that the log-average miss rate of the algorithm reached 27.6%. Shojaei et al. [9] used transfer component analysis and maximum independent domain in pedestrian target detection. The experimental results on the dataset of INRIA showed that the pedestrian target detection algorithm with domain adaptation had less classification error. Wang et al. [10] designed an algorithm using image fusion and deep learning to improve the performance of unmanned aerial vehicles for detecting pedestrians on the ground in low-illumination environments and verified the excellent performance of the algorithm through experiments.

The previous text is a review of some studies related to pedestrian target detection, and different researchers have used different approaches to identify and detect pedestrian targets. In general, the basic principle of these pedestrian target recognition and detection methods is to extract pedestrian features from images and recognize them based on the extracted features. However, the extracted image features under different scales were not fully considered in the above-mentioned studies; therefore, in this paper, the image features at different scales were utilized.

This paper studied intelligent algorithms for pedestrian target detection on roads. This paper was written in the following structure. The abstract starts with a general statement of the full paper. The introduction gives a brief overview of the related literature. Then, the pedestrian target detection algorithm is described, including the traditional SVM algorithm and the improved Faster RCNN algorithm. Then, the simulation experiment is described. In the experiment, the SVM algorithm, traditional Faster RCNN, and improved Faster RCNN algorithms were compared. The final conclusion summarizes the results of this paper. The contribution of this paper is to optimize the Faster RCNN algorithm for pedestrian target detection, so that it can make full use of the convolutional feature maps at different scales, providing an effective reference for accurate and fast detection of pedestrian targets on the road. The limitation of this paper is that the types of images used in the training of the algorithm were not comprehensive enough, so the richness of the types of images required for algorithm training will be increased to improve the generalizability of the algorithm in the future.

## II. AUTOMATIC DETECTION ALGORITHM FOR PEDESTRIANS

### A. Traditional Pedestrian Target Detection Algorithm

A video consists of multi-frame images, so the detection of pedestrians in the video can be considered as fast detection of pedestrians in the image. The traditional pedestrian target detection algorithm extracts features from the images in the candidate frames, uses a classification algorithm to classify and

identify the images in the candidate frames according to the extracted features, and takes the candidate frames judged to be pedestrians as the output. Its specific steps are illustrated in Fig. 1.

1) An input image is pre-processed. A plural number of candidate boxes are added to the image. The size of the candidate boxes is determined according to the actual application.

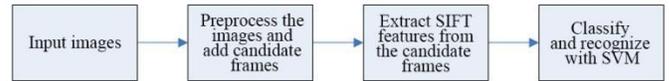


Fig. 1. Traditional pedestrian target detection algorithm.

2) Scale-invariant feature transform (SIFT) feature extraction is performed on the image in the candidate frame. The extraction of SIFT features requires the Difference of Gaussian (DoG) operator [11] to construct a Gaussian difference pyramid. The calculation formula of the DoG operator is:

$$D(x, y, \sigma) = (G(x, y, k\sigma) - G(x, y, \sigma)) \times I(x, y), \quad (1)$$

where  $D(x, y, \sigma)$  denotes the DoG operator, whose scale factor is  $\sigma$ ,  $G(x, y, \sigma)$  and  $G(x, y, k\sigma)$  are the Gaussian filter functions, whose scale factor is  $\sigma$  and adjacent to  $\sigma$ , respectively, and  $I(x, y)$  is the original image. Then, the local extreme points of the image in every scale in the Gaussian difference pyramid composed of DoG operators are searched. The gradient histogram is constructed by choosing the appropriate neighborhood range with the extreme point in every level of the pyramid as the center [12]. Eventually, the histograms corresponding to the main direction of every extreme point and the direction greater than 80% of the gradient peak of the main direction are merged as the SIFT feature.

3) The SIFT features of the collected image sample are separated into a training group and a test group. The training group is used to train and fit the SVM to get the classification function. After the training, the SVM classification function is used to determine whether the image in the candidate frame is a pedestrian according to the SIFT features of the image sample.

### B. Pedestrian Target Detection Algorithm using Convolutional Neural Network

In the traditional pedestrian detection algorithm described in the previous text, the features of the image are firstly extracted before recognition by the SVM, which simply means that the extraction of the image features and the recognition of the image are independent of each other. Moreover, the extracted SIFT features are statistical, which are difficult to fully reflect the features of the image and will affect the detection accuracy of the algorithm.

A CNN, as a deep learning algorithm [13], can extract local features of images by convolutional kernels, and the plural features obtained from the plural convolutional kernels can be

combined into global features, taking into account the global and local. The Faster RCNN algorithm is a CNN algorithm for detecting pedestrian targets. It first extracts the convolution feature map of the image through the convolutional and pooling structures of a conventional CNN and obtains the candidate target frame from the map through a regional proposal network (RPN) [14]. After the convolutional features in the candidate target frame are pooled by region of interest (ROI), whether the target frame is a pedestrian is determined in the fully connected layer, and regressive calculation is also performed on the target frame that is judged as a pedestrian in the fully connected layer to get the coordinates of the target frame in the original image.

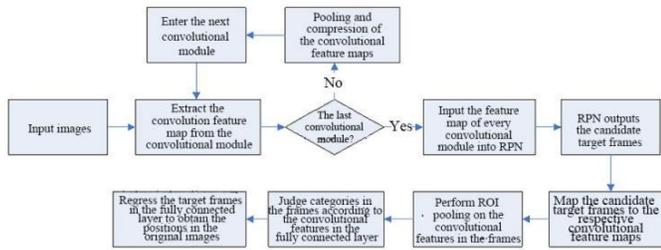


Fig. 2. Pedestrian detection process of the optimized Faster RCNN algorithm.

The convolutional and pooling structures of the CNN in the Faster RCNN algorithm will produce convolutional feature maps of different scales. In this paper, in order to make good use of the convolutional features of different scales to improve pedestrian detection accuracy, some improvements are made to the Faster RCNN algorithm. The optimized detection process is presented in Fig. 2.

- 1) A pre-processed image is input into the input layer.
- 2) Convolutional features are extracted in the convolutional module.
- 3) Whether the convolutional module is the last convolutional module in the conventional CNN structure is determined. If not, the convolutional features are pooled and compressed. The compressed convolutional feature map is input into the next convolutional module for the operation in step 2; if it is, the convolutional feature map of the last convolutional layer obtained in every convolutional module is input into the RPN.
- 4) The candidate target frame is obtained after calculation in RPN: In this structure, the pixel points in the feature map are regarded as anchor points, and every anchor point generates nine candidate frames with three scales and three length-width ratios with itself as the center [15]. The candidate frame score is calculated according to the convolution features in the candidate frame; the higher the score, the higher the probability of the candidate frame being the target frame. Some candidate frames with high probability are selected and mapped to the original image according to the ratio of the feature map where the candidate frame is located to the original image, and the candidate frames that are beyond the boundary of the original image are deleted. Some candidate frames with high probability are chosen from the remaining candidate frames again as the output of RPN.

5) The candidate target frames calculated by RPN are mapped to the respective convolutional feature maps to which they belong, i.e., the candidate target frames are mapped to the feature maps from which they are obtained.

6) ROI pooling operation [16] is performed on the convolutional features in the candidate target frame. The convolutional map in the target frame is divided into regions according to the required size for ROI pooling, and every region is processed by max-pooling. For example, if the feature map with a size of  $9 \times 9$  in the target frame needs to be compressed into a size of  $3 \times 3$ , the feature map in the target frame is divided into regions in a size of  $3 \times 3$ , every region is processed by max-pooling, and the result is taken as the value of the corresponding region.

7) The convolutional features processed by ROI compression are input into the fully connected layer to determine whether they are pedestrians, and the position of the target frame in the original image is calculated [17].

The improvement of the optimized Faster RCNN algorithm compared to the traditional Faster RCNN algorithm is that instead of using only the convolutional feature map given by the last convolutional module, convolutional feature maps of different scales in the previous convolutional module are used, making full use of the convolutional features of different scales.

### III. SIMULATION EXPERIMENTS

#### A. Experimental Setup

The algorithm in Fig. 3 has five convolutional modules. Convolutional modules 1 and 2 both have two convolutional layers, and there are 32 convolutional kernels in a size of  $3 \times 3$  in every layer [18]. Convolutional modules 3, 4 and 5 all have three convolutional layers, and there are 64 convolutional kernels in a size of  $3 \times 3$  in every layer. Convolutional modules 1~4 have 1 pooling layer after every module, every pooling layer uses a pooling frame in a size of  $2 \times 2$ , and the mean-pooling is used in the pooling frame. The RPN module is a fully convolutional structure. Convolutional feature maps obtained from convolutional layers 2, 4, 7, 10, and 13 are all used to calculate the candidate target frames in the RPN. The ROI pooling layer compresses the convolutional features in the candidate target frames, and the compressed size is  $6 \times 6$ . The fully connected layer recognizes the category of convolutional features after ROI pooling to determine whether the image in the target frame is a pedestrian. Moreover, the regressive calculation is conducted on the candidate target frame to obtain the coordinates of the target frame in the original image.

The images collected by the author were used as the dataset for the simulation experiment. The images came from a variety of scenes, not limited to traffic intersections. After preliminary removing images with too blurred pedestrians and too dark backgrounds, 15,210 images were left, and the scenes included traffic intersections, parks, supermarkets, subway stations, neighborhoods, etc. Sixty percent of the images were used as the training samples, and the remaining 40% as the test samples.

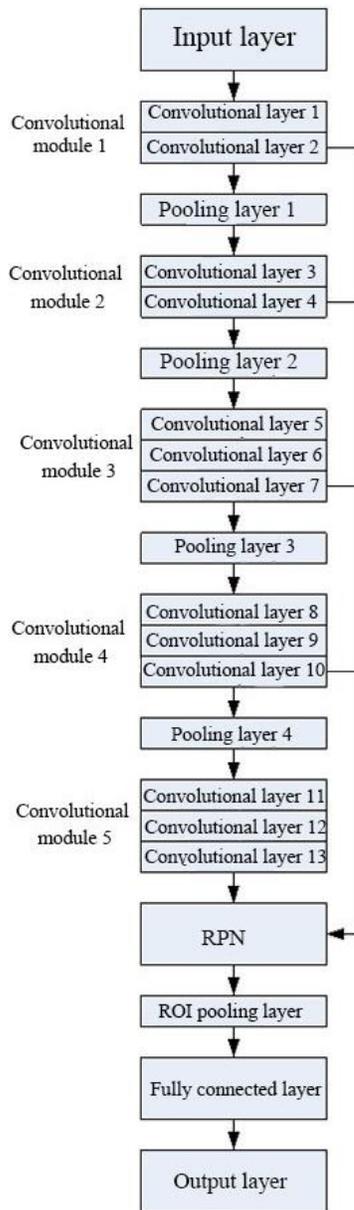


Fig. 3. Basic structure of the improved faster RCNN algorithm.

In the simulation experiment, two detection algorithms, the SVM algorithm and the traditional Faster RCNN algorithm, were also tested to further verify the performance of the improved Faster RCNN algorithm. The SVM algorithm identified pedestrians in the images with SIFT features, and the size of the target frame used for extracting SIFT features was  $6 \times 6$ . The basic structure of the traditional Faster RCNN algorithm was similar to that of the optimized Faster RCNN algorithm, and their only difference was that the convolutional feature maps in convolutional layers 2, 4, 7, and 10 were not input into the RPN.

### B. Evaluation Criteria

Target detection for pedestrians is a binary classification problem, i.e., to determine whether the target in an image target frame is a pedestrian, so the performance of the detection algorithm can be evaluated using a confusion matrix [19], as

shown in Table I. The detection precision and recall rate are calculated using the following equations:

$$\begin{cases} P = \frac{TP}{TP + FP} \\ R = \frac{TP}{TP + FN} \end{cases}, \quad (2)$$

where  $P$  is the precision and  $R$  is the recall rate. In addition, the detection speed of the pedestrian target detection algorithm is also quite important. Here, frame per second (FPS) was used to measure the detection speed of the algorithm, i.e., the number of images detected per unit time.

In addition to the above evaluation criteria, the author also used Intersection over Union (IoU) to measure the target frame positioning accuracy of the algorithm. The calculation formula of IoU is:

$$IoU = \frac{DR \cap GT}{DR \cup GT}, \quad (3)$$

where  $IoU$  stands for the degree of target frame overlap,  $DR$  denotes the target frame predicted by the algorithm, and  $GT$  denotes the actual target frame in the image.

### C. Experimental Results

The SVM algorithm and traditional Faster RCNN algorithm were compared with the optimized Faster RCNN algorithm. Due to the limitation of space, only some of the detection results are displayed. Fig. 4 shows the pedestrian target detection results of three algorithms for the same image. It was seen from Fig. 4 that the SVM algorithm marked the relatively obvious pedestrians in the image but missed smaller pedestrians, and moreover, it identified two pedestrians as one pedestrian among the marked pedestrians, so it was not very effective in recognizing pedestrian targets overall. In the result of the conventional Faster RCNN algorithm, more pedestrians were detected than in the SVM algorithm, and the two pedestrians that overlap in the picture were also distinguished, but it also missed smaller pedestrian targets. The improved Faster RCNN algorithm not only detected and distinguished relatively significant pedestrians but also detected smaller pedestrian's targets, so its detection performance was the best.

Fig. 5 shows the precision and recall rate of the SVM, traditional Faster RCNN, and optimized Faster RCNN algorithm for the test set. The precision of the SVM algorithm for pedestrian target detection was 75.3%, and the recall rate was 73.8%; the precision of the traditional Faster RCNN algorithm for pedestrian target detection was 86.7%, and the recall rate was 85.7%; the precision of the improved Faster RCNN algorithm had a precision of 96.6% and a recall rate of 95.4%.

TABLE I. CONFUSION MATRIX

	Pedestrian actually	Background actually
Judged as pedestrian	$TP$	$FP$
Judgment as background	$FN$	$TN$



Fig. 4. Partial detection results of three pedestrian target detection algorithms.

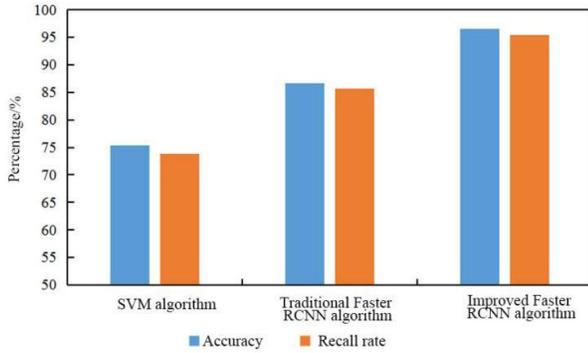


Fig. 5. Detection performance of the three algorithms.

Fig. 6 shows the detection speed of SVM, traditional Faster RCNN, and improved Faster RCNN algorithms for the test set. The detection speed of the SVM, traditional Faster RCNN, and improved Faster RCNN algorithms for pedestrian targets was 10.36 FPS, 21.33 FPS, and 33.51 FPS, respectively. It was seen from Fig. 6 that the SVM algorithm had the lowest detection speed, the traditional RCNN algorithm had a detection speed higher than the SVM algorithm, and the improved Faster RCNN algorithm had a detection speed higher than the traditional RCNN algorithm.

Fig. 7 shows the target frame localization accuracy of the three pedestrian target detection algorithms. The IoU of the target frame of the SVM, traditional Faster RCNN, and improved Faster RCNN algorithms was 67.3%, 79.8%, and 93.4%, respectively. It was observed in Fig. 7 that the target frame obtained by the SVM algorithm in the process of pedestrian detection had the lowest degree of overlap with the actual target frame, the degree of overlap between the target frame obtained by the traditional Faster RCNN and the actual target frame was higher than that of the SVM algorithm, and the degree of overlap between the target frame calculated by the improved algorithm and the actual target frame was higher than that of the traditional Faster RCNN algorithm.

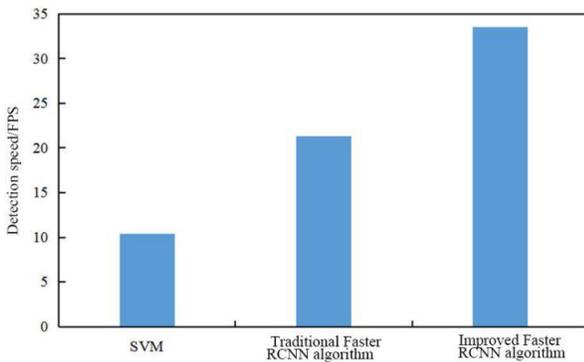


Fig. 6. Detection speed of three pedestrian target detection algorithms.

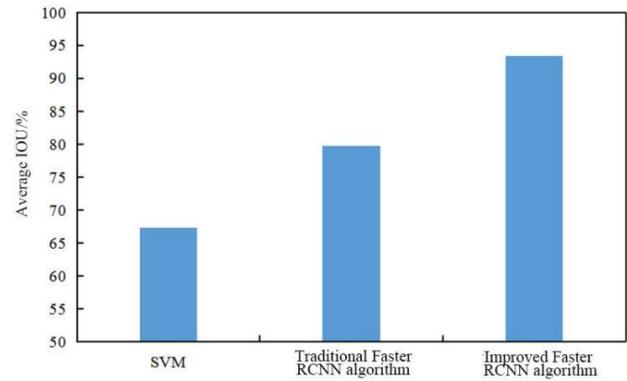


Fig. 7. Target frame localization accuracy of three pedestrian target detection algorithms.

Based on the comparison results of detection accuracy and speed among the three pedestrian target detection algorithms, it was found that the improved Faster RCNN algorithm had the best detection performance, followed by the traditional Faster RCNN algorithm, and the SVM algorithm performed the poorest. The reasons are shown below. The SVM algorithm extracted SIFT features when detecting pedestrian targets, which were statistical features that could not fully reflect the features in the image, so it was difficult to distinguish overlapped pedestrians in the image in the detection process, and the fixed size of the target frame also made it inflexible to distinguish smaller pedestrian targets. The traditional Faster RCNN algorithm used the convolutional structure of a CNN to extract the local and global features of the image, so it performed better than the SVM algorithm in recognition, but only the convolutional features of the last convolutional layer were used in the calculation of the candidate target frame. Even if the target frames with three length-width ratios were used, it was difficult to effectively use the multi-scale features, so the traditional algorithm missed the pedestrians. In the improved Faster RCNN algorithm, the convolutional features of different scales were fully utilized in the calculation of candidate frames, so it effectively recognized smaller pedestrian targets in the images. Moreover, the SVM algorithm used fixed-size target frames for recognition and identified all the target frames, leading to a low detection speed and decreased localization accuracy of the computed target frames; the traditional Faster RCNN algorithm used RPN to pre-compute candidate frames, which reduced the repetitiveness of target frame selection, and different scales of candidate frames improved the localization accuracy of small targets; the improved Faster RCNN algorithm used convolutional features of different scales in the computation of target candidate frames and further improved the localization accuracy of target frames for small pedestrians by using target frames of different scales.

#### IV. CONCLUSION

This paper compared the SVM, traditional Faster RCNN, and improved Faster RCNN algorithms in simulation experiments after improving the traditional Faster RCNN algorithm. The experimental results are shown below. (1) The detection results of some images showed that the improved Faster RCNN algorithm effectively distinguished the pedestrians in the image as well as the background and also

achieved better detection results when facing small target pedestrians in the image compared to the other two algorithms. (2) In terms of detection accuracy for pedestrians in images, the detection accuracy and recall rate of the improved Faster RCNN was 75.3% and 73.8%, respectively; the traditional Faster RCNN algorithm was 86.7% and 85.7%, respectively; the SVM algorithm was 96.6% and 95.4%, respectively. (3) In terms of the detection speed, the detection speed of the SVM algorithm was 10.36 FPS, the traditional Faster RCNN algorithm was 21.33 FPS, and the improved Faster RCNN algorithm was 33.51 FPS. (4) In terms of the localization frame accuracy, the IoU of the target frame of the SVM algorithm was 67.3%, the IoU of the traditional Faster RCNN algorithm was 79.8%, and the IoU of the improved Faster RCNN algorithm was 93.4%.

#### REFERENCES

- [1] C. B. Murthy, M. F. Hashmi, G. Muhammad, and S. A. AlQahtani, "YOLOv2PD: An Efficient Pedestrian Detection Algorithm Using Improved YOLOv2 Model," *Comput. Mater. Contin.*, pp. 3015-3031, January 2021.
- [2] Z. Xu, W. Zhao, L. Peng, and J. Chen, "Research on Pedestrian Detection Algorithm Based on Deep Learning," *J. Phys. Conf. Ser.*, vol. 1646, pp. 1-6, September 2020.
- [3] H. Xu, M. Guo, N. Nedjah, J. Zhang, and P. Li, "Vehicle and Pedestrian Detection Algorithm Based on Lightweight YOLOv3-Promote and Semi-Precision Acceleration," *IEEE T. Intell. Transp.*, vol. 23, pp. 19760-19771, January 2022.
- [4] H. Xia, J. Ma, J. Ou, X. Lv, and C. Bai, "Pedestrian detection algorithm based on multi-scale feature extraction and attention feature fusion," *Digit. Signal Process.* vol. 121, pp. 1-13, November 2022.
- [5] S. S. Liu, "Self-adaptive scale pedestrian detection algorithm based on deep residual network," *Int. J. Intell. Comput.*, vol. 12, pp. 318-332, August 2019.
- [6] J. Yang, W. Y. He, T. L. Zhang, C. L. Zhang, L. Zeng, and B. F. Nan, "Research on subway pedestrian detection algorithms based on SSD model," *IET Intell. Transp. Sy.*, vol. 14, pp. 1491-1496, November 2020.
- [7] Y. Zhang, K. Guo, W. Guo, J. Zhang, and Y. Li, "Pedestrian crossing detection based on HOG and SVM," *J. Cyber Secur.*, vol. 2, pp. 79-88, January 2021.
- [8] D. Pei, M. Jing, H. Liu, L. Jiang, and F. Sun, "A fast retinanet fusion framework for multi-spectral pedestrian detection," *Infrared Phys. Techn.*, vol. 105, pp. 1-8, January 2020.
- [9] G. Shojaei and F. Razzazi, "Semi-supervised domain adaptation for pedestrian detection in video surveillance based on maximum independence assumption," *Int. J. Multimed. Inf. R.*, vol. 8, pp. 241-252, December 2019.
- [10] C. Wang, D. Luo, Y. Liu, B. Xu, and Y. Zhou, "Near-surface pedestrian detection method based on deep learning for UAVs in low illumination environments," *Opt. Eng.*, vol. 61, pp. 1-19, February 2022.
- [11] J. Wang, C. Zhao, Z. Huo, Y. Qiao, and H. Sima, "High quality proposal feature generation for crowded pedestrian detection," *Pattern Recogn.*, vol. 128, pp. 1-10, February 2022.
- [12] H. Zhou, and G. Yu, "Research on Fast Pedestrian Detection Algorithm Based on Autoencoding Neural Network and AdaBoost," *Complexity*, vol. 2021, pp. 1-17, March 2021.
- [13] Z. J. Wang, Y. Q. Zhao, and C. L. Zhao, "Improved MSER Pedestrian Detection Algorithm based on TOF Camera," *J. Phys. Conf. Ser.*, vol. 1576, pp. 1-6, June 2020.
- [14] S. Zhai, S. Dong, D. Shang, and S. Wang, "An Improved Faster R-CNN Pedestrian Detection Algorithm Based on Feature Fusion and Context Analysis," *IEEE Access*, vol. 8, pp. 138117-138128, January 2020.
- [15] S. Y. Cho, J. H. Lee, and C. G. Park, "A Zero-Velocity Detection Algorithm Robust to Various Gait Types for Pedestrian Inertial Navigation," *IEEE Sens. J.*, vol. 22, pp. 4916-4931, March 2021.
- [16] J. Ren, C. Niu, and J. Han, "An IF-RCNN Algorithm for Pedestrian Detection in Pedestrian Tunnels," *IEEE Access*, vol. 8, pp. 165335-165343, January 2020.
- [17] G. Li, C. Zong, G. Liu, and T. Zhu, "Application of Convolutional Neural Network (CNN)-AdaBoost Algorithm in Pedestrian Detection," *Sensor. Mater.*, vol. 32, pp. 1997-2006, June 2020.
- [18] D. Liu, S. Gao, W. Chi, and D. Fan, "Pedestrian detection algorithm based on improved SSD," *Int. J. Comput. Appl. T.*, vol. 65, pp. 25-35, January 2021.
- [19] B. Wang, "Research on Pedestrian Detection Algorithm Based on Image," *J. Phys. Conf. Ser.*, vol. 1345, pp. 1-12, November 2019.

# Enhanced Multi-Verse Optimizer (TMVO) and Applying it in Test Data Generation for Path Testing

Mohammad Hashem Ryalat<sup>1\*</sup>, Hussam N. Fakhouri<sup>2</sup>, Jamal Zraqou<sup>3</sup>, Faten Hamad<sup>4</sup>, Mamon S. Alzboun<sup>5</sup>, Ahmad K. Al hwaitat<sup>6</sup>

Department of Computer Science, Al-Balqa Applied University, Salt, Jordan<sup>\*1</sup>,

Department of Data Science & Artificial Intelligence, University of Petra, Amman, Jordan<sup>2</sup>.

Department of Virtual and Augmented Reality, University of Petra, Amman, Jordan<sup>3</sup>

Department of Information Studies, Sultan Qaboos University, Muscat, Oman<sup>4</sup>

Department of Curricula and Instruction, Al al-Bayt University, Mafraq, Jordan<sup>5</sup>

Department of Computer Science, University of Jordan, Amman, Jordan<sup>6</sup>

**Abstract**—Data testing is a vital part of the software development process, and there are various approaches available to improve the exploration of all possible software code paths. This study introduces two contributions. Firstly, an improved version of the Multi-verse Optimizer called Testing Multi-Verse Optimizer (TMVO) is proposed, which takes into account the movement of the swarm and the mean of the two best solutions in the universe. The particles move towards the optimal solution by using a mean-based algorithm model, which guarantees efficient exploration and exploitation. Secondly, TMVO is applied to automatically develop test cases for structural data testing, particularly path testing. Instead of automating the entire testing process, the focus is on centralizing automated procedures for collecting testing data. Automation for generating testing data is becoming increasingly popular due to the high cost of manual data generation. To evaluate the effectiveness of TMVO, it was tested on various well-known functions as well as five programs that presented unique challenges in testing. The test results indicated that TMVO performed better than the original MVO algorithm on the majority of the tested functions.

**Keywords**—MVO; optimization; testing; swarm intelligence; multi-verse optimizer

## I. INTRODUCTION

The term "optimization" describes the process of identifying the most optimal search solutions that are likely to resolve a particular issue. There is more than one conventional and meta-heuristic optimization strategy available. The standard techniques are gradient-based and have a faster execution time than convergence. On the other hand, these methods are not applicable to multimodal functions that are neither differentiable nor predictable. Thus, this technique does not allow for the discovery of the global optimal solution. Due to the fact that they start with only one point, it gets trapped at the local optimal value. There are many other search strategies that can be used to solve this problem; however, most of them require additional assistance that is based on exponential time, which makes them more time-consuming [1]. As a result, meta-heuristic optimization approaches have become the most widely used approach. Meanwhile, intelligent algorithms are increasingly used in the development of applications, testing, and the making of business decisions in today's world [2][3].

The use of meta-heuristics has been increasingly widespread over the past two decades. Computer researchers in a wide variety of domains are familiar with meta-heuristic techniques such as the Genetic Algorithm, multi-verse, and Particle Swarm Optimization, amongst others. Because of its ease of use, adaptability, and absence of approaches requiring derivation, meta-heuristic has garnered a lot of attention in recent years [4] [5]. Techniques for testing software include both black-box and white box testing. In the black-box method of software testing, the tester is only privy to the system's architecture. He or she is not privy to any information regarding the program's internal design and does not have access to the source code. Its purpose is to guarantee that the system accepts all of the necessary inputs in the way that was described and produces results that are accurate. White box testing, also known as structural testing, focuses on investigating the internal logic and structure of the source code being tested. During the structural test, each possible code path will be checked for a predetermined set of test information inputs. It is very important to select a diverse control flow way to test since there are a large number of paths for test succession, and performing the tests in succession can be difficult. Finding connections between system components, choosing those paths, creating test data for every path, and assessing test results are only a few examples of the many problem viewing paths involved in software testing [6].

The white box test criteria for software testing, such as branch coverage, focus on the process of locating a group of test cases that increases the likelihood of error discovery. Within the context of this approach, an experiment will serve as the indication that triggers the calling of the test routines with specific input group values. After that, those drivers will make a comparison between the output and the one that was relied upon. Utilizing known inputs that can be put to use but will ultimately prove to be impossible, allowing for an infinite supply of them. As a result, the primary focus of automated software testing is on the process of naturally locating the smallest set of inputs in order to broaden the scope of the test criteria [7]. When it comes to the process of developing test cases for critical path coverage testing, the concept of linear coded sequence is absolutely necessary. It is possible that the productivity of the development of all of these important paths

can be increased, and at the same time, it will be appropriate to create test cases with the assistance of a variety of testing tools so that those tested programs can be investigated. This can be done through all of the important paths [8].

It has been shown by the "No Free Lunch" theorem [9] that metaheuristics do not always succeed in solving optimization issues. Results show that metaheuristic optimization works well for one type of optimization issue but not another. For the aforementioned causes, it is important to create a more efficient optimization metaheuristic algorithm [10], [11].

The primary contribution of this study is the introduction of an improved version of the Multi-verse Optimizer, named the Testing Multi-Verse Optimizer (TMVO). Instead of focusing on a specific region, TMVO takes into account the mobility of the swarm and the average of the two best solutions across the universe. A mean-based algorithm model is employed to guide particle movement towards the optimal solution. TMVO's proposed movement equations enable effective space exploration and utilization, and also address the issue of poor convergence, providing an additional benefit by escaping local minima.

The second contribution of this study involves the application of TMVO algorithm, an enhanced swarm intelligence metaheuristic, to address the issue of single objective optimization in the automated generation of test cases for structural data testing, particularly path testing. Rather than automating the entire testing process, TMVO focuses on centralizing automated procedures for collecting testing data. The proposed TMVO achieves this goal by directing the swarm based on the past performance of the top three solutions discovered by the swarm. The population search history is also utilized to provide an alternative answer, which is the mean of the three best spots identified so far, thus improving the particles' ability to explore the space. This results in more opportunities for the swarm particles to be discovered and utilized, thereby increasing the likelihood of achieving a global optimum while avoiding a local minimum challenge. To overcome these challenges, the direction of particle flow is switched with each cycle.

Due to the absence of a universally applicable metaheuristic that can be used to address all optimization issues, and the fact that no metaheuristic has proven to be effective for solving all identified optimization problems, many swarm intelligence studies have focused on optimizing specific systems.

Route testing is a methodology for testing software that involves a search of the program domain for test cases that, when combined with the code, will cause the program to follow a specified path. Path testing is an optimization issue with no unique solution due to the unlimited number of possible pathways in a program. Consequently, it is only realistic to pick a fraction of these paths for testing. If the pathways to be tested have been clearly described and an adequate fitness function has been constructed, then TMVO might be used for this purpose. In this work, a test case is treated as a representative of a generation, with the chosen target route serving as the endpoint toward which the algorithm is directed.

This study aimed to address one of the most well-known problems in software testing by proposing an improved swarm intelligence metaheuristic method, called TMVO, to resolve the route testing problem. The TMVO method was created to address the aforementioned issues and proposed a better route for the swarm particles to follow, improving the movement strategy of a swarm of particles. To evaluate the algorithm's efficacy, a battery of benchmark functions was used, and its exploitation, exploration, global optimal solution, and best path-finding abilities were tested across these three domains. The results were compared to those of a popular metaheuristic technique, and several indicators, both visual and statistical, were used to assess the quality of the output. The proposed enhanced technique successfully solved the single-objective optimization issue in software testing.

The following goals have been set for this research; the first goal is to propose an improved MVO optimization method by averaging the best places in the search space, which is informed by the past motion of the particles. The second goal is to use the superior movement approach to increase the efficacy of swarm movement in path testing and test data collection. The third goal is to use the created metaheuristic to address the MVO premature, to converge problem and the local optima entrapment problem. The fourth goal is to compare the proposed enhanced method to existing optimization algorithms through empirical testing using standard benchmark functions and testing software.

In this work, the Testing Multi-Verse Optimizer (TMVO) is presented as an improved Multi-verse Optimizer. Instead than focusing on a single place, TMVO considers the swarm's mobility and the mean of the two best options in the universe. Using a mean-based algorithm model that has been suggested, particles will migrate toward the ideal solution. The recommended movement equations of TMVO ensure the effectiveness of space exploration and utilization. In addition to resolving the issue of poor convergence, it also escapes the local minimum.

This study makes a contribution through enhancing MVO in solving the problem of path testing by enhancing the test data generation. It also provides a comprehensive analysis of the algorithm's movement strategy, equations, pseudo-code, and parameters. When it comes to solving software testing issues, the algorithm offers a more effective path testing method for getting to best tested path. TMVO has been evaluated and validated in comparison to a number of well-established functions. In addition to this, it provides a solution for a problem involving a single optimization problem in software testing.

The remaining part of this study is organized as follows. The related works are reviewed in more detail in Section II. The methodology including different types of software testing, and the path coverage test is described in Section III. In Section IV, the experimental results and discussion are presented where Section V concludes this study.

## II. RELATED WORK

The Multi-Verse optimizer, often known as the MVO, was first suggested to be developed by Mirjalili and colleagues

[12]. They came up with an original algorithm that was inspired by nature and gave it the name Multi-Verse Optimizer (MVO). The white hole, the black hole, and the wormhole are the three natural phenomena that serve as the inspiration for this algorithm's design. The demand for these models arises from the requirement to independently carry out exploration, exploitation, and vicinity search. Biswas [13] presented an ant colony optimization (ACO)-based method that produces groups of ideal pathways and ranks them in order of preference. In addition, using these methodologies leads in the grouping of test data inside the area so that similarity may be used as input for the paths that are constructed. The proposed methods ensure comprehensive software coverage with little duplication of effort. In [14], the authors employed an approach dubbed "propagation error" to analyze the growth of defects. Through the development of test cases, we are able to activate seed faults and provoke associated potential issues. The testing procedure involves triggering and correlating these flaws. Clever algorithms are used in this method, with the aim of permanently designing test cases to disperse data about seed flaws. All faults and related defects that were before invisible are now easily discernible thanks to propagation routes.

Aspect oriented programming (AOP) is recommended by Jain et al. [15], [16] as a method for crawling into program modules without modifying their source code and component in order to investigate regions where faults are suspected to exist. AOP execution places an emphasis on making use of system cut points. In addition to this, it includes crucial code at each execution point for the purpose of testing. To improve the effectiveness of conventional random testing and random partition testing approaches, some researches suggested using Dynamic Random Testing, also known as DRT. The DRT is presented as a potential further improvement to the testing's viability. In order to decide on those upgrades for a testing profile that is more reasonable, it is necessary to have access to additional historical testing data along with an estimation of the rate at which defects are identified for each subdomain in real time, for example. This exemplifies one instance of the symbolization that the Java-based DSU system provides. In this approach, system tests that were developed for both older and newer versions of the program can be updated, and it purposefully tests whether or not an incremental upgrade can result in a failed test.

Testing software is widely regarded as an effective strategy for ensuring the quality of software in both the academic and commercial settings. The quality of the test data has an effect on the testing process and is also an essential component in determining how well software is tested. As a natural part of the software development life cycle, software testing may be carried out either automatically or manually as a matter of course. Both approaches have their advantages and disadvantages. The creation of test data is the initial step in the software testing process. In the testing process, there are a few various procedures that need to be carried out. These procedures include the development of test data, the prioritizing of test cases, and the reduction of test cases. The initialization of the test data is the method that is the most difficult aspect of testing in these methods. According to [17],

there would be a variety of sub-tasks amongst test cases, test appropriateness, and test data [18].

Test cases are the conditions that are going to be set, and the analyzer is going to use those to determine whether or not the specified function fits in suitably. The gathering of test cases will ensure that the test is suitable. Test data are a special sort of data that is used for evaluating different software applications. They can be easily recognized from other types of data. In addition, it will serve as the feed for the system's input. It is possible that this will serve as the principal test for the data or the field validations for any software applications. Creating test data for very simple programs is not a tough undertaking. On the other hand, producing the data for extensive initiatives might be challenging [19]. There is a wide variety of software available that can be used to generate test data [20], including intelligent test data generators, test data generators that use path oriented principles, and test data generators that use goal oriented. Creating test data would involve the use of several methodologies, such as UML diagrams; nevertheless, the development of test data would be dependent on graphical user interfaces. The coverage-based testing methodology, which consists of a collection of conditions that absolutely need to fulfill all of the prerequisites, could be used to generate test data [21]. A wide variety of coverage strategies, including branch coverage, function coverage, and statement coverage, are all viable options.

However, there is no assurance that the flaws in the test data will be uncovered by every converge method. The offered strategies leverage objective function for test data creation. The test data that are generated as a result of the objective function give the best possible possibility for defect detection. The space and path disparity functions are the goal functions. In order to get the space disparity, we need to first measure the distance that separates each of the test suites. Next, we need to calculate the path disparity by working backwards from the branch condition through the control flow graph [22], [23]. Because product testing must take into account both the long term and the cost-benefit analysis, extensive testing may not be carried out. Since a wide variety of methods and resources are used to automate the processes [24], it's possible that the use of such mechanizations for testing has become essential as of late. Successful testing requires the identification of code routes, the creation of a test data suit for those paths, a testing procedure on the Software Under Test (SUT) using the data, an evaluation of the results, and the production of quality models.

Successful testing would examine as many test cases as possible that are similar to those already performed. As an added cost-cutting measure, it is important to prioritize paths with the expectation that the majority of errors will be found in the preliminary phases of the process, and to identify appropriate paths and test data from among the many possible options. Path testing is a very useful technique for finding bugs in software components [25], [26].

### III. METHODOLOGY

In this section, we describe the procedures and techniques employed to carry out the study, including data analysis, and statistical methods. The research design and settings are also discussed in detail. This section provides a detailed account of

the methods used to answer the research questions and provides a clear understanding of how the research was conducted.

#### A. Types of Testing

It is of vital importance to clarify here the main types of testing since testing is used in this study to test the research hypothesis. The testing of software can be divided into two categories: static testing and dynamic testing.

In static software testing, the reviewer completes code reviews by walking through hypothetical inputs to the SUT while outwardly accompanying the real program flow. Static testing is a type of software testing. This method requires the reviewer to invest their time, and the reviewer themselves need to be an expert as well as possess the necessary skills to evaluate the code. It is possible to specify from these variables the paths that might not be executable. This is made possible by the enhancements to static testing that let the code be symbolically evaluated. This is done by gathering distinct paths and variables regarding code execution. This methodology could be used to aggregate these variables in order to provide a demand solver with the information it needs to decide which routes and paths were previously infeasible.

When performing dynamic testing, the SUT code may actually be executed using the test inputs that have been provided. The observed behaviors of the SUT are compared to its typical behaviors, and the test is either successful or unsuccessful depending on whether the observed behaviors match the technique that is relied upon to conduct the test. There are two different kinds of testing that may be done on dynamic systems: black-box testing and white box testing. The outcome of an output defect is what is understood to be a software defect [27].

In black-box testing, the system is evaluated without the tester having any prior knowledge of the system's underlying architecture. In black-box testing, the individual performing the testing does not have access to the program's source code. He or she needs knowledge regarding the modeling of the framework. In this section, the tester generally connects with the software through the user interface by providing inputs and testing outputs. However, the tester is not expected to have any prior knowledge regarding how to operate input. The accuracy of software objectives is checked for throughout the black-box testing process. These objectives can be tested using the inputs and outputs domain. This demonstrates that the program in question has both an input and an output; results from output failures are regarded to be software flaws [28].

Testing with a black-box can be used to identify problems with data structures, error functions, and interfaces. Black-box eliminates system techniques. It detects errors that are caused by faults in the software in order to find out what the problem is with the output. It is possible to use it to identify incorrect functions, which produced undesirable output at executed, inaccurate conditions. This is due to the fact that incorrect functions generate inaccurate outputs anytime they are put into action.

Testing procedures that provide information regarding the internal specification and design of the system are referred to

as white-box testing. It is not unusual for this to be referred to as structural testing. It includes testing for anything to do with program logic, including testing for loops, testing conditions, and testing based on data flow. Even if there is only an incomplete software definition, this will assist in the discovery of flaws. The goal of white box testing is to ensure that each possible path in software has been explored by the test cases.

White box testers have access to the system's source code and are therefore familiar with its architecture. The tester begins by analyzing the source code, then uses the knowledge from the source code to generate a variety of test cases, and finally, particular code routes are utilized in order to achieve a desired amount of code coverage [29]. It is guaranteed by the test cases that each of the program's independent pathways has been followed at least once. Each internal data structure would be tested to ensure the system's dependability. Each loop is run until it reaches its boundaries while staying within its operational constraints. White-box testing is a technique that can be utilized by software engineers in the process of designing test cases. This technique involves practicing distinct paths within a module, practicing legitimate true and false decisions, executing loops at their limits and inside their operational limits, and practicing inner data structures to guarantee that they are correct. It would appear that test cases need to be modified whenever implementation is altered. In this article, we have simply utilized the black-box testing approach to evaluate the functionality of two separate lines based on different test cases utilizing BVA and Robustness testing. White box testing, on the other hand, covers testing the majority of the program's code. Changing the requirements under test conditions will help identify typographical problems [30].

#### B. Path Coverage Test

The testing technique known as "coverage basic path testing" refers to testing strategies that are designed to cover the fundamental path of the software. The test target is the fundamental flow of the program when it is executed using this method. After gathering test information for the program input space, taking those test data into consideration as input, and then eventually running the program, it carries out the fundamental path by running the program and executing it. The participation of the fundamental routes group is required in order to carry out the genuine testing technique. The following is a list of features that are shared by all fundamental paths: 1) Each and every path in the program is completely autonomous; 2) Each and every edge in the program is accessible; and 3) Any paths in the program that do not have a position with the path set can potentially be achieved through the use of paths linear operation in the fundamental path set. The fault propagation path is a way that will show the advancement of defects where mistakes originate in software nodes; they may gradually propagate on different nodes. This method will be referred to as the fault propagation path. During the procedure that is used to repair errors that have already been created, past errors will be used to determine which paths have the greatest potential for error propagation. This will help correct errors that have already been made. Inaccurate historical data will be used as a source of this knowledge, and it will be used to define these routes.

The MVO algorithm uses the expansion rate as the determining factor for the value of the function for each and every search. In addition, every particle in the search zone has a similar appearance to an elected solution as well as a variable in an elected solution. Greater expansion rates result in greater and lower possibilities of the existence of those hypothesized white holes and black holes, respectively. These higher expansion rates also bring search agents or universes with higher rates to transfer items through those white gaps. White holes are recommended as a result of reduced inflation rates, which also reduce the expansion rates that should be used to transport items into black holes. As a result, the probability of black holes is increased, and white holes are offered as a result. Wormholes, disregard the flatland rates; they would be the explanation for the arbitrary sending of the object to the best universes. The MVO algorithm contains a wheel choice component that can be used for scientific demonstrations of white holes and black holes, as well as the return of objects to the search area. The search agents are arranged in each iteration according to their expansion rates, and once a search agent is chosen, it must be assigned a white hole. These various characteristics of the universes are supported by MVO. It makes use of wormholes in order to transport irregular things through the search region, and it does so by exploiting those wormholes. These wormholes randomly switch the positions of those objects in the search region, preventing them from claiming their expansion rates in any scenario. Wormhole connections have to be helped along between our reality and the finest possible universe.

### C. The proposed Multi-Verse Optimizer (TMVO)

This sub-section introduces the proposed TMVO, including the algorithm steps, pseudo-code, the strategy, TMVO's operations, and its parameters, and theoretical conclusion.

TMVO is a stochastic swarm optimization algorithm with a revolutionary exploration and exploitation movement approach for locating optimum solutions to optimization problems. TMVO is based on enhancing MVO movement strategy by taking the top three solutions in the swarm for the automatic development of test cases for structural data testing, particularly path testing. Since the original MVO algorithm lacked the ability to effectively cover both the exploration and exploitation stages of the search process, the TMVO algorithm was developed to solve this problem. In addition, TMVO addresses the premature convergence issue that arises with certain implementations of the MVO algorithm. TMVO algorithm advises focusing exploration and exploitation efforts on the following points: White holes would be a higher amount of time on make in the universes for secondary expansion rates, which they transmit items on distant universes. This is because white holes consume an inordinate quantity of matter and energy. In addition to this, assist them in improving their rates of expansion. Black holes would appear in universes with low expansion rates, and as a result, they provide a higher probability of items being accepted from other universes. This is because low expansion rates result in more compact universes. This adds another layer to the possibility of claiming an increasing inflation rate for universes that have a lower expansion rate. White and black hole tunnels have a tendency to transport from worlds the objects with rising expansion rates

to the folks with low expansion rates; in this method, the general inflation rate concerning known universes will be moved forward across the span from those repetitions. Wormholes have a propensity to appear in any universe at random, regardless of the expansion pace, or something along those lines due to the many properties of. Through all of the repetitions, the universe remains preserved. If there is a sudden shift, white/black hole tunnels need universes, which will lead to an inquiry of the search space. Unanticipated progressions are also helpful in determining the ideal local solidity. Random wormholes re-expansion of the variables from variables of the universes around the finest result gained in this way in those course about iterations, thus ensuring that exploitation is performed around those the overwhelming majority guaranteeing area of the search region. WEP Adaptive values expansion will concentrate exploitation by using an optimization procedure. This is because the occurrence of wormholes in universes is a likelihood. TDR Adaptive values reduce the journey variable distance near the best universe. This is a method that expands the precision of a local search through iterations. The joining of those indicated by the algorithm is ensured by checking the exploitation of local search comparative of the amount derived from the number of iterations.

The following are the main steps involved in TMVO:

The first step, which is named initialization, involves initially populating the algorithm's parameters with random values. Second, the suggested motion equations will be used to iteratively improve upon these initial best guess answers.

The second step: the algorithm's designated equations are utilized to progressively enhance the outcomes until a stopping criterion is met.

The third step: The algorithm's optimal solution is determined by balancing the values of the goal function and comparing the resultant comparisons.

The pseudo-code for the TMVO algorithm is displayed in Fig. 1.

1. Define the set of all universes,  $U$ .
2. Define the set of all portfolio weights,  $w$ .
3. Define the set of all groups,  $G$ .
4. Initialize the current universe,  $u$ , and the current portfolio weights,  $w$ .
5. Evaluate the performance of the current portfolio,  $P$ , in the current universe,  $u$ .
6. For each group,  $g$ , in  $G$ :
  - a. Select a subset of universes,  $U'$ , from  $U$  that belong to group  $g$ .
  - b. For each universe,  $u'$ , in  $U'$ :
    - i. Calculate the portfolio weights,  $w'$ , that maximize the expected return in universe  $u'$ .
    - ii. Evaluate the performance of the portfolio,  $P'$ , in universe  $u'$  using weights  $w'$ .
  - c. Select the universe,  $u^*$ , and the corresponding portfolio weights,  $w^*$ , that result in the highest performance in the subset of universes.
7. Select the group,  $g^*$ , and the corresponding universe,  $u^{**}$ , and portfolio weights,  $w^{**}$ , that result in the highest overall performance.
8. Update the current universe and portfolio weights to  $u^{**}$  and  $w^{**}$ , respectively.
9. Go to step 5 and repeat the process.

Fig. 1. TMVO pseudocode.

An exploration phase and an exploitation phase are separated by a population-based method, as we saw in the previous section. For MVO space exploration, it has employ white hole and black hole ideas. On the other hand, the wormholes help MVO make better use of the search spaces. We treat every possible answer as if it were its own world, with each variable representing a different type of thing that may be found in that universe. The value of the fitness function is used to determine the inflation rate that is applied to each solution. As time is a standard concept in both cosmology and multi-universe theory, we employ it throughout this study rather than iteration.

However as in MVO, the TMVO universes are optimized using the following criteria: If inflation rates are high enough, white holes are almost guaranteed to form. Black holes are less likely to form with greater inflation rates. Third, items in universes with a higher inflation rate are more likely to be sucked into white holes.

The number of items that enter the universe via black holes is larger in universes with a lower inflation rate. No matter the pace of inflation, things in all worlds may eventually make their way through wormholes to the best universe at random.

TMVO Pseudocode show that the normalized inflation rate serves as a roulette wheel for selecting and determining white holes. The likelihood of sending things through white hole or black hole tunnels increases as the inflation rate decreases. When solving the maximizing problems, -NI must be replaced with NI. Since the universes must swap things and experience sudden changes in order to traverse the search space, the exploration can be ensured using this approach.

The previously mentioned method allows for unabated object exchange across worlds. We assume that each universe is equipped with wormholes that allow its things to randomly travel through space, allowing for the preservation of cosmological variety while also allowing for the possibility of exploitation. Wormholes are capable of altering the objects of universes at random, regardless of their inflation rates. We assume that wormhole tunnels are constantly built between a universe and the best universe generated so far in order to supply local modifications for each universe and have a high likelihood of enhancing the inflation rate utilizing wormholes.

The suggested methods have varying degrees of computing complexity, which are determined by the number of iterations, the number of universes, the roulette wheel mechanism, and the universe sorting mechanism. Every iteration includes the process of sorting the universe, and we use the Quicksort algorithm, which, in the best case scenario, has a complexity of  $O(n \log n)$ , and in the worst case scenario, has a complexity of  $O(n^2)$ . The selection from the roulette wheel is carried out for each variable in each universe throughout the iterations, and its complexity ranges from  $O(n)$  to  $O(\log n)$ , depending on the implementation.

The followings are some observations that are concluded in order to gain an understanding of how the suggested algorithm could, in theory, have the ability to solve optimization problems:

- White holes are more likely to form in universes that have high inflation rates since this increases the likelihood that they will be able to transmit things to other universes and help those universes increase their inflation rates.
- Black holes are more likely to emerge in worlds with low inflation rates because these holes have a larger likelihood of receiving things from other universes since inflation rates are lower. This once more raises the possibility of increasing inflation rates for those universes that now have low inflation rates.
- The general or average inflation rate of all universes steadily improves over the course of the iteration process, as white and black hole tunnels tend to carry objects from universes with high inflation rates to those with low inflation rates.
- Because wormholes tend to form at random in any world, independent of the inflation rate, the variety of universes may be kept intact throughout the course of several iterations of the simulation.
- Wormhole and black hole tunnels need the sudden transformation of universes, which ensures the thorough investigation of the search space.
- Sudden shifts are helpful in resolving local optimalities that have stagnated.
- During the iterative process, wormholes randomly reposition some of the variables in the universes surrounding the best solution gained thus far. This facilitates exploitation all over the most promising area of the search space.
- The existence probability of wormholes in universes is gradually increased when adaptive WEP settings are used. As a result, the process of optimization places a strong emphasis on exploitation.
- To enhance the precision of local search during iterations, adaptive TDR values are used to decrease the variable's traveling distance around the best universe.
- By placing a greater emphasis on exploitation and local search in relation to the number of iterations, the suggested algorithm's convergence is ensured.

#### IV. EXPERIMENTAL RESULTS AND ANALYSIS

##### A. Evaluation of TMVO over the Benchmark Functions

To test the performance of TMVO, experiments had been run over well-known benchmark functions that represent unimodal and multi-modal functions that have been used by many researchers [31][32][33].

The cost functions of the benchmark unimodal function (F1-F7) are displayed in Table I, and those for the multimodal functions (F8-F14) are displayed in Table II. In order to get reliable statistical findings, the experiment needs to be carried out  $n$  times before any meaningful conclusions can be drawn about the performance of meta-heuristic algorithms. Each run needs to be carried out until  $m$  numbers of iterations have been

completed, and this is for the purpose of verifying if the algorithm is stable. In most cases, the statistical and output metrics, such as the average, the standard deviation, as well as the minimum and maximum values, of the best solution in the most recent iteration are measured and registered for comparison studies of the algorithms. For the purposes of acquiring, recording, and verifying the outcomes of the TMVO algorithm, the exact same process and experimental approach have been adhered to throughout. In addition to computing the error, it is important to determine how much the findings deviate from the ideal value.

TABLE I. UNIMODAL FUNCTIONS MATHEMATICAL FORMULATION (F1-F7)

No.	Formula
F1	$f1(x) = \sum_{i=1}^n x_i^2$
F2	$f2(x) = \sum_{i=1}^n  x_i  + \prod_{i=1}^n  x_i $
F3	$f3(x) = \sum_{i=1}^n \left( \sum_{j=1}^n x_j \right)^2$
F4	$f4(x) = \max i\{ x_i , 1 \leq i \leq n\}$
F5	$f5(x) = \sum_{i=1}^{n-1} [100(x_{i+1} - x_i^2)^2 + (x_i - 1)^2]$
F6	$f6(x) = \sum_{i=1}^n (x_i + 0.5)^2$
F7	$f7(x) = \sum_{i=1}^n ix_i^4 + \text{random}(0,1)$

TABLE II. MULTIMODAL BASIC FUNCTIONS (F8-F14)

No.	Formula
F8	$f8(x) = \sum_{i=1}^n -x_i \sin \sqrt{ x_i } * \sum_{i=1}^n ix_i^4 * \text{random}(0,1) *$
F9	$f9(x) = \sum_{i=1}^n [x_i^2 - 10 \cos(2\pi x_i) + 10]$
F10	$f10(x) = -20 \exp \left( -0.2 \sqrt{\frac{1}{n} \sum_{i=1}^n x_i^2} \right) - \exp \left( \frac{1}{n} \sum_{i=1}^n \cos(2\pi x_i) \right) + 20 + e$
F11	$f11(x) = \frac{1}{4000} \sum_{i=1}^n x_i^2 - \prod_{i=1}^n \cos \left( \frac{x_i}{\sqrt{i}} \right) + 1$
12	$f12(x) = \frac{\pi}{n} \left\{ 10 \sin(\pi y_1) + \sum_{i=1}^{n-1} (y_i - 1)^2 [1 + 10 \sin^2(\pi y_{i+1})] + (y_n - 1)^2 \right\} + \sum_{i=1}^n u(x_i, 10, 100, 4)$ $y_i = 1 + \frac{x_i + 1}{4} u(x_i, a, k, m) = \begin{cases} k(x_i - a)^m & x_i > a \\ 0 & -a < x_i < a \\ k(-x_i - a)^m & x_i < -a \end{cases}$
F13	$f13(x) = 0.1 \left\{ \sin^2 + \sum_{i=1}^n (x_i - 1)^2 [1 + \sin^2(3\pi x_i + 1)] + (x_n - 1)^2 [1 + \sin^2(2\pi x_n)] \right\} + \sum_{i=1}^n u(x_i, 5, 100, 4)$
F14	$f14(x) = -\sum_{i=1}^n \sin(x_i) \cdot \left( \sin \left( \frac{ix_i^2}{\pi} \right) \right)^{2m}, m=10$

The average, on the other hand, compare the overall performance of the method. All of the statistical analyses that were carried out allow us to establish beyond a reasonable doubt that the results were not the product of random chance. In each of the algorithms, the population size was set at fifty, and the maximum number of iterations was set at one thousand. It is important to keep in mind, however, that the maximum number of iterations and the number of particles (possible solutions, for example) should be determined by experimentation when dealing with situations that occur in real life.

It is necessary to conduct tests a total of n times if one wishes to achieve reliable statistical findings from meta-heuristic algorithms. In addition, for the purpose of validating the consistency of the method, each iteration must be carried out until the m<sup>th</sup> time. In order to create TMVO, report on its performance, and then validate its efficacy, the identical experimental process was carried out.

The effectiveness of the TMVO algorithm that was proposed has been assessed. It has been proved that there is a set of statistical measurements that includes the average, the standard deviation, the minimum, the maximum, and the error measurement. These measurements have been determined through the process of experimentation throughout the course of the twenty-three benchmark functions shown in the Tables (1-2).

The primary regulating parameters of these algorithms, the number of search particles and the maximum iteration, have been set to the values of 50 and 1000 respectively so that a fair comparison can be made between them. To achieve the highest possible level of performance, the settings for the various governing parameters of each algorithm are taken from the most recent version of the source code. Each of the algorithms is executed fifty times on each of the test functions, and the outcomes of these simulations are presented later in this study. It is important to note that the results of the algorithms are standardized in the range [0, 1] by employing the min-max normalization so that their performances may be compared across a variety of test functions.

We have evaluated the performance of TMVO on a set of well-known benchmark functions utilized by many researchers to measure the performance of optimization algorithms.

The benchmark sets for multimodal hybrid functions are categorized from function 15 to function 23 and the mathematical formulations for hybrid composition functions are shown in Table III.

The Lower Bound (LB), Upper Bound (UB), dimension (Dim), and F<sub>min</sub> of the benchmark-evaluated functions are displayed in Table IV.

TABLE III. MULTIMODAL HYBRID FUNCTIONS (F8-F14)

Multimodal Functions Formula (F15- F23)
$f15(x) = \sum_{i=1}^{11} \left[ a_i - \frac{x_1(b_i^2 + b_i x_2)}{b_i^2 + b_i x_3 x_4} \right]$

$f16(x) = 4x_1^2 - 2.1x_1^4 + \frac{1}{3}x_1^6 + x_1x_2 - 4x_2^2 + 4x_2^4$
$f17(X) = \left(x_2 - \frac{5.1}{4\pi^2}x_1^2 + \frac{5}{\pi}x_1 - 6\right)^2 + 10\left(1 - \frac{1}{8\pi}\right)\cos X_1 + 10$
$f18(x) = [1 + (x_1 + x_2 + 1)^2(19 - 14x_1 + 3x_1^2 - 14x_2 + 6x_1x_2 + 3x_2^2)]x[30 + (2x_1 - 3x_2)^2x(18 - 32x_1 + 12x_1^2 + 48x_2 - 36x_1x_2 + 27x_2^2)]$
$f19(x) = -\sum_{i=1}^4 C_i \exp\left(-\sum_{j=1}^3 a_{ij}(x_j - p_{ij})^2\right)$
$f20(x) = -\sum_{i=1}^4 C_i \exp\left(-\sum_{j=1}^6 a_{ij}(x_j - p_{ij})^2\right)$
$f21(x) = -\sum_{i=1}^5 [(X - a_i)(X - a_i)^T + C_i]^{-1}$
$f22(x) = -\sum_{i=1}^7 [(X - a_i)(X - a_i)^T + C_i]^{-1}$
$f23(x) = -\sum_{i=1}^{10} [(X - a_i)(X - a_i)^T + C_i]^{-1}$

TABLE IV. PARAMETERS OF THE EVALUATED FUNCTIONS F1-F23

Function	Dim	LB	UB	F <sub>min</sub>
F1	30	-100.00	100.00	Zero
F2	30	-10.00	10.00	Zero
F3	30	-100.00	100.00	Zero
F4	30	-100.00	100.00	Zero
F5	30	-30.00	30.00	Zero
F6	30	-2400.00	2400.00	Zero
F7	30	-1.28	1.28	Zero
F8	30	-500.00	100.00	418.9829x5
F9	30	-5.12	5.12	Zero
F10	30	-32.00	32.00	Zero
F11	30	-600.00	600.00	Zero
F12	30	-2400.00	2400.00	Zero
F13	30	-2400.00	2400.00	Zero
F14	2	-5.00	5.00	1
F15	4	-5.00	5.00	0.00030
F16	2	-5.00	5.00	1.0316
F17	2	[-5,0]	[10,15]	0.398
F18	2	-2.00	2.00	3.00
F19	3	0.00	1.00	-3.86
F20	6	-5.00	5.00	-3.32
F21	4	0.00	10.00	-10.1532
F22	4	0.00	10.00	-10.4028
F23	4	0.00	10.00	-10.5363

TABLE V. COMPARISON BETWEEN THE TMVO AND MVO IN TERMS OF MEAN FITNESS VALUE

F#	Mean of TMVO	Mean of MVO	TMVO vs MVO
F 1	1.89618	20.80666	Better
F 2	1.482	5.2544	Better
F 3	8.9748	234.41	Better
F 4	1.4287	3.9419	Better
F 5	67.3715	843.6999	Better
F 6	6.5543	22.9003	Better
F 7	0.070841	0.86756	Better
F 8	-4864.7373	-7388.2747	No
F 9	25.6411	163.2599	Better
F 10	2.8175	4.5803	Better
F 11	0.91005	1.6356	Better
F 12	2.5573	2.4075	No
F 13	0.077985	0.2351	Better
F 14	0.998	0.998	Equal
F 15	0.00044931	0.00058644	Better
F 16	-1.0316	-1.0316	Equal
F 17	0.3978	0.3978	Equal
F 18	3	3	Equal
F 19	-3.8628	-3.8627	Better
F20	-3.3214	-3.201	Better
F21	-10.0464	-2.6068	Better
F22	-10.3418	-9.8605	Better
F23	-5.1928	-5.1885	Better

The comparison between the proposed TMVO algorithm and the MVO algorithm in terms of mean fitness value is tabulated in Table V. Comparing TMVO algorithm with MVO over the tested functions F1-F23 showed that TMVO has very competitive results. In the unimodal functions (F1-F7) the TMVO has shown better results and outperformed MVO over all the seven functions. Regarding the multi-modal functions (F8-F12), TMVO was also achieved better mean fitness values than MVO except F8. Moreover, the proposed algorithm is competitive over the expanded multi-modal functions (F13, F14). The results have shown that when testing the algorithm TMVO over the multi-modal hybrid functions (F15-F23), the TMVO outperformed MVO in most cases and achieved the same fitness value in three cases.

The proposed TMVO achieved better fitness values in most cases due to the fact that TMVO offers additional exploration points inside the search space. TMVO takes the two best possible solutions and utilizes them to find a new solution at each iteration. It drives closer and closer to the global optimum by updating the current particle location to the position that is optimal between these two points.

*B. Evaluation of TMVO in Test Data Generation for Path Testing*

The experimental results testing is carried out on five benchmark programs, which are presented in Table VI. The fitness value is a numerical number that represents individual

quality in comparison with the existing local solution in order to seek for the optimum local solution that has the least amount of fitness value possible. The option that results in the lowest overall fitness value will be the one that we consider to be the most viable solution. The fitness value is computed by applying Korel's route distance relation to each variable. The fitness path distance is calculated by adding up each variable's fitness value at each point along the path. To start, a series of random test instances are generated so that the process can begin. Utilizing points that were picked at random allows for the improvement of the existing solution. Perform a calculation to determine the fitness value of each potential solution. Each swarm is assigned a fitness value, and then each swarm searches for the local minimum value within the search zone to see whether a higher value can be found. If we can, the new value is saved, and the old value is replaced with it. Arrange candidate solutions in order of increasing fitness, beginning with the best. The onlooker phase begins with the most optimal solution to fitness. If the termination requirements are deemed to be complete, an onlooker local search will be issued; otherwise, it will be used to improve candidate solution fitness. In case that the phase is completed without satisfying the finishing conditions, the phase to replace sources that have reached the maximum number of tries will be initiated.

We utilized the five variables in Program1, which are (x, y, z, j, k). First, if  $j - 80 \geq 0$ , the distance at variable j will be zero; if variable  $k - 70 \geq 0$ , the distance at variable k will be zero; if variable  $x - 60 \geq 0$ , the distance at variable y - 50 = 0; this is the Korel branch distance relation. Specific details and results including the fitness value are tabulated in Table VII. It is of vital importance to correctly interpret the values in Table VII. The letters A, B, C, D, and E represent the Korel's route branch distance of the variables j, k, c, y, and z respectively.

TABLE VI. BENCHMARK PROGRAMS USED AS CASE STUDIES

Program1	Program2	Program3
If (j >=80) {..... } Else if (k >= 70) {..... } Else If (x >=60) {..... } Else If (y>=50) {..... } Else If (z>=25) {..... }	while (j >=75) {..... } while (k >= 65) {..... } while (x >=55) {..... } while (y>=45) {..... } while (z>=35) {..... }	If (j >=60) {..... } Else If (k >= 80) {..... } Else If (x >=55) {..... } if (y>=25) {..... } while (z>=45) {..... }
Program4	Program5	
If (j >=57) {..... } If (k >= 68) {..... } if (x >=34) {..... }	If (j >=45) {..... } Else If (k >= 30) {..... } while (x >=40) {..... } Else If (y>=35) {..... }	

TABLE VII. KOREL'S ROUTE BRANCH DISTANCES OF THE VARIABLES J, K, C, Y, AND Z ALONG WITH THE FITNESS VALUES (PROGRAM 1)

#	j	k	x	y	z	A	B	C	D	E	Fit.
1	91	50	75	100	54	11	0	15	50	29	105
2	89	64	84	68	66	9	0	24	18	41	92
3	81	87	71	82	87	1	17	11	32	62	123
4	62	72	89	52	99	0	2	29	2	74	107
5	84	56	72	86	79	4	0	12	36	54	106
6	70	91	84	92	59	0	21	24	42	34	121
7	77	67	71	72	88	0	0	11	22	63	96
8	76	97	61	84	65	0	27	1	34	40	102
9	53	65	72	79	85	0	0	12	29	60	101
10	90	56	80	80	70	10	0	20	30	45	105
11	96	66	62	53	75	16	0	2	3	50	71
12	99	63	61	88	85	19	0	1	38	60	118
13	55	87	90	55	85	0	17	30	5	60	112
14	78	95	60	72	93	0	25	0	22	68	115
15	68	98	63	93	93	0	28	3	43	68	142
16	76	97	85	69	51	0	27	25	19	26	97
17	99	57	79	84	68	19	0	19	34	43	115
18	59	92	85	75	84	0	22	25	25	59	131
19	100	93	100	59	82	20	23	40	9	57	149
20	59	67	72	94	76	0	0	12	44	51	107
21	67	87	62	58	59	0	17	2	8	34	61
22	100	80	76	90	69	20	10	16	40	44	130
23	66	78	95	58	82	0	8	35	8	57	108
24	50	54	54	66	86	0	0	0	16	61	77
25	63	78	89	98	51	0	8	29	48	26	111

The following equation is utilized to determine the fitness value that is used for the path of Program1. This value, which is the sum of the distances that were indicated before, is computed as follows:

$$F = (J - 80) + (K - 70) + (X - 60) + (Y - 50) + (Z - 25) \quad (1)$$

In Program2, we utilized the five variables (x, y, z, j, k). If the first variable (j) has a distance of zero, then the second variable (k) also has a distance of zero. (if (k) - 65 >= 0), the third variable (x) has a distance of zero. if (x) - 55 >= 0, the fourth variable (y) has a distance of zero if (y) - 45 = 0, and the fifth variable (z) has a distance of zero if (z) - 35. The Korel's route branch distances of the variables in Program2 and the fitness values are displayed in Table VIII.

Eq. (2) has been used to get the fitness value that should be utilized for the path of program 2, which is 54. This value represents the sum of the distances that were indicated earlier.

$$F = (J - 75) + (K - 65) + (X - 55) + (Y - 45) + (Z - 35) \quad (2)$$

TABLE VIII. KOREL'S ROUTE BRANCH DISTANCES OF THE VARIABLES J, K, X, Y, AND Z ALONG WITH THE FITNESS VALUES (PROGRAM 2)

#	j	k	x	y	z	A	B	C	D	E	Fit.
1	86	84	52	94	78	11	19	0	49	43	122
2	92	57	59	86	100	17	0	4	41	65	127
3	74	54	62	63	83	0	0	7	18	48	73
4	86	57	94	75	71	11	0	39	30	36	116
5	66	96	83	52	81	0	31	28	7	46	112
6	70	98	69	58	95	0	33	14	13	60	120
7	81	88	92	92	80	6	23	37	47	45	158
8	53	69	86	91	100	0	4	31	46	65	146
9	53	81	64	74	68	0	16	9	29	33	87
10	89	79	53	82	94	14	14	0	37	59	124
11	51	58	88	97	84	0	0	33	52	49	134
12	78	66	99	54	84	3	1	44	9	49	106
13	86	61	66	73	89	11	0	11	28	54	104
14	100	57	98	51	57	25	0	43	6	22	96
15	77	68	78	63	100	2	3	23	18	65	111
16	91	80	92	68	57	16	15	37	23	22	113
17	78	76	82	62	68	3	11	27	17	33	91
18	86	65	60	67	56	11	0	5	22	21	59
19	72	73	72	67	72	0	8	17	22	37	84
20	87	73	100	84	67	12	8	45	39	32	136
21	61	64	92	61	50	0	0	37	16	15	68
22	95	50	69	78	68	20	0	14	33	33	100
23	62	53	52	51	83	0	0	0	6	48	54
24	52	91	54	60	73	0	26	0	15	38	79
25	94	96	83	80	100	19	31	28	35	65	178

In Program3, the five variables (x,y,z,j,k) are also employed to evaluate the proposed algorithm. The Korel branch distance relation states that if the value of the first variable, j, is greater than or equal to 60, then the value of the second variable, k, is greater than or equal to 80. If the value of the third variable, x, is greater than or equal to 45, the value of the fourth variable, y, is less than or equal to 75, and the value of the fifth variable, z, is 45. Table IX tabulates the outcomes when applying the TMVO over Program3. The symbols (A, B, C, D, E) represent Korel's Route Branch Distances of the variables (j, k, x, y, z) respectively.

TABLE IX. KOREL'S ROUTE BRANCH DISTANCES OF THE VARIABLES J, K, X, Y, AND Z ALONG WITH THE FITNESS VALUES (PROGRAM 3)

#	j	k	x	y	z	A	B	C	D	E	Fit.
1	58	85	77	72	75	0	5	22	47	30	104
2	94	70	83	83	91	34	0	28	58	46	166
3	83	90	88	76	93	23	10	33	51	48	165
4	70	55	87	96	100	10	0	32	71	55	168
5	86	94	64	60	93	26	14	9	35	48	132
6	77	90	60	79	72	17	10	5	54	27	113
7	73	51	64	73	68	13	0	9	48	23	93
8	66	50	86	83	64	6	0	31	58	19	114
9	79	75	53	61	76	19	0	0	36	31	86

10	97	63	76	54	91	37	0	21	29	46	133
11	87	86	67	84	93	27	6	12	59	48	152
12	75	83	65	71	53	15	3	10	46	8	82
13	80	91	97	64	95	20	11	42	39	50	162
14	90	72	81	62	63	30	0	26	37	18	111
15	71	100	54	81	64	11	20	0	56	19	106
16	79	81	84	66	61	19	1	29	41	16	106
17	71	55	71	64	93	11	0	16	39	48	114
18	66	88	78	75	71	6	8	23	50	26	113
19	52	88	80	63	82	0	8	25	38	37	108
20	79	76	69	71	68	19	0	14	46	23	102
21	75	84	61	88	50	15	4	6	63	5	93
22	56	67	83	95	72	0	0	28	70	27	125
23	79	98	69	61	60	19	18	14	36	15	102
24	53	100	53	84	57	0	20	0	59	12	91
25	58	80	75	65	95	0	0	20	40	50	110

The fitness value that was used for the path of program3 was 82, which is the sum of the distances that were indicated earlier and is computed using Equation 3 as follows

$$F = (J - 60) + (K - 80) + (X - 55) + (Y - 45) + (Z - 25) \quad (3)$$

TABLE X. KOREL'S ROUTE BRANCH DISTANCES OF THE VARIABLES X, Y, AND Z ALONG WITH THE FITNESS VALUES (PROGRAM4)

#	x	y	z	C	D	E	Fit.
1	100	52	61	43	0	27	70
2	64	97	77	7	29	43	79
3	66	99	66	9	31	32	72
4	92	85	80	35	17	46	98
5	89	91	62	32	23	28	83
6	67	88	65	10	20	31	61
7	80	50	71	23	0	37	60
8	93	63	65	36	0	31	67
9	86	72	53	29	4	19	52
10	99	98	80	42	30	46	118
11	56	56	72	0	0	38	38
12	53	82	54	0	14	20	34
13	95	82	55	38	14	21	73
14	56	70	96	0	2	62	64
15	93	55	76	36	0	42	78
16	56	80	78	0	12	44	56
17	55	56	60	0	0	26	26
18	100	95	51	43	27	17	87
19	52	80	55	0	12	21	33
20	53	100	94	0	32	60	92
21	93	61	64	36	0	30	66
22	96	58	82	39	0	48	87
23	90	77	60	33	9	26	68
24	70	64	75	13	0	41	54
25	88	84	97	31	16	63	110

In Program4, we employed three variables (x, y, z). If  $j - 60 \geq 0$ ,  $k - 80 \geq 0$ , and  $x - 45 \geq 0$ , then the distance between the first and third variables is zero, as predicted by the Korel branch distance relation. Refer to Table X. The symbols (C, D, E) represent Korel's Route Branch Distances of the variables (x, y, z) respectively.

Using Equation 4, we can determine that the fitness value for path of program4 is 33, which is the total of the distances we determined before.

$$F = (J - 57) + (K - 68) + (X - 34) \quad (4)$$

Four variables were employed which are j, k, x, and y in Program5. In the Korel branch distance relation, if the value of the first variable, j, is zero, then the value of the second, k, is also zero, and so on. If the value of the third variable, x, is also zero, then the value of the fourth one, y, is also zero. Table XI tabulates 25 different cases along with their fitness values.

Path 5 of Program5 uses a fitness value of 39, which is the total of the distances discussed before. The fitness value is calculated according to Equation5.

$$F = (J - 45) + (K - 30) + (X - 40) + (Y - 35) \quad (5)$$

TABLE XI. KOREL'S ROUTE BRANCH DISTANCES OF THE VARIABLES J, K,X, AND Y ALONG WITH THE FITNESS VALUES (PROGRAM5)

#	j	k	x	y	A	B	C	D	Fit.
1	93	95	73	52	36	27	39	0	102
2	64	64	93	64	7	0	59	0	66
3	89	50	90	68	32	0	56	0	88
4	74	73	96	79	17	5	62	10	94
5	89	53	70	50	32	0	36	0	68
6	54	76	98	88	0	8	64	19	91
7	55	97	73	76	0	29	39	7	75
8	98	61	81	91	41	0	47	22	110
9	99	84	88	52	42	16	54	0	112
10	80	73	62	77	23	5	28	8	64
11	54	64	89	82	0	0	55	13	68
12	55	69	59	98	0	1	25	29	55
13	88	52	61	92	31	0	27	23	81
14	50	59	99	64	0	0	65	0	65
15	71	58	55	83	14	0	21	14	49
16	64	100	89	92	7	32	55	23	117
17	98	56	84	57	41	0	50	0	91
18	91	74	97	52	34	6	63	0	103
19	64	75	55	73	7	7	21	4	39
20	89	83	56	95	32	15	22	26	95
21	75	85	64	56	18	17	30	0	65
22	66	75	51	89	9	7	17	20	53
23	80	52	79	94	23	0	45	25	93
24	99	70	90	89	42	2	56	20	120
25	70	86	66	52	13	18	32	0	63

## V. CONCLUSION

In this study, Testing Multi-Verse Optimizer (TMVO), an improved Multi-Verse Optimizer, is presented. However, rather than focusing on a single place, TMVO considers the swarm's mobility and the mean of the two best solutions in the universe. Using a recently suggested mean-based algorithm model, particles will progress toward the ideal solution. TMVO's recommended movement equations ensure efficient space exploration and utilization. In addition, it eliminates the problem of low convergence and escapes the local minimum. TMVO has been applied for the generation of test data for software structural testing, specifically route testing, that takes use of the Multi-Verse optimization algorithm. The proposed algorithm has been exhaustively tested through the creation of test data for the path coverage criteria and its subsequent application to a set of test programs. Additionally, five distinct programs and codes have been utilized in order to complete this evaluation. The results showed that the algorithm was successful in finding the best tested path for the test data, which led to an improvement in performance. The performance of TMVO is tested over several well-known functions. The results have shown that TMVO outperform original MVO algorithm over most of the tested functions.

However, this study presented two contributions. Firstly, an improved version of the Multi-verse Optimizer called Testing Multi-Verse Optimizer (TMVO) was proposed, which considered the movement of the swarm and the mean of the two best solutions in the universe. The particles moved towards the optimal solution by using a mean-based algorithm model, which guaranteed efficient exploration and exploitation. Secondly, TMVO was applied to develop test cases for structural data testing, specifically path testing, in an automated manner. Instead of automating the entire testing process, the focus was on centralizing automated procedures for collecting testing data. Automation for generating testing data was becoming increasingly popular due to the high cost of manual data generation. To evaluate the effectiveness of TMVO, it was tested on various well-known functions as well as five programs that presented unique challenges in testing. The test results indicated that TMVO outperformed the original MVO algorithm on the majority of the tested functions.

Despite the success of TMVO, there are still several areas where the algorithm can be further developed and tested. This includes algorithmic parameter tuning where most optimization algorithms have several tuning parameters that need to be set for optimal performance. Future research can explore automated parameter tuning techniques such as machine learning algorithms to improve the performance of TMVO. In addition to that, testing TMVO on large-scale problems where researchers can focus on testing TMVO on large-scale optimization problems and analyzing its scalability and efficiency.

## REFERENCES

- [1] Shukri, S. E., Al-Sayyed, R., Hudaib, A., & Mirjalili, S. (2021). Enhanced multi-verse optimizer for task scheduling in cloud computing environments. *Expert Systems with Applications*, 168, 114230.
- [2] Jamunaa, D., Mahanti, G. K., & Hasoon, F. N. (2022). Multi-verse optimization algorithm for optimal synthesis of phase-only reconfigurable linear array of mutually coupled parallel half-wavelength

- dipole antennas placed at finite distances from the ground plane. *Scientia Iranica. Transaction D, Computer Science & Engineering, Electrical*, 29(4), 1915-1924.
- [3] Hamad, F., Al-Aamr, R., Jabbar, S. A., & Fakhuri, H. (2021). Business intelligence in academic libraries in Jordan: Opportunities and challenges. *IFLA Journal*, 47(1), 37-50.
- [4] Yadav, M., & Mishra, A. (2023). An enhanced ordinal optimization with lower scheduling overhead based novel approach for task scheduling in cloud computing environment. *Journal of Cloud Computing*, 12(1), 1-14.
- [5] Ryalat, M. H., Dorgham, O., Tedmori, S., Al-Rahamneh, Z., Al-Najdawi, N., & Mirjalili, S. (2022). Harris hawks optimization for COVID-19 diagnosis based on multi-threshold image segmentation. *Neural Computing and Applications*, 1-19.
- [6] Song, R., Zeng, X., & Han, R. (2020). An improved multi-verse optimizer algorithm for multi-source allocation problem. *International Journal of Innovative Computing, Information and Control*, 16(6), 1845-1862.
- [7] Aljarah, I., Mafarja, M., Heidari, A. A., Faris, H., & Mirjalili, S. (2020). Multi-verse optimizer: theory, literature review, and application in data clustering. *Nature-inspired optimizers: theories, literature reviews and applications*, 123-141.
- [8] Pan, R., Bagherzadeh, M., Ghaleb, T. A., & Briand, L. (2022). Test case selection and prioritization using machine learning: a systematic literature review. *Empirical Software Engineering*, 27(2), 29.
- [9] Schaeffer, R., Khona, M., & Fiete, I. (2022). No free lunch from deep learning in neuroscience: A case study through models of the entorhinal-hippocampal circuit. *bioRxiv*, 2022-08.
- [10] Fakhouri, H. N., Hamad, F., & Alawamrah, A. (2022). Success history intelligent optimizer. *The Journal of Supercomputing*, 1-42.
- [11] Aldabbas, H., Asad, M., Ryalat, M. H., Malik, K. R., & Qureshi, M. Z. A. (2019). Data augmentation to stabilize image caption generation models in deep learning. *Int J Adv Comput Sci Appl*, 10(10), 571-9.
- [12] Mirjalili, S., Mirjalili, S. M., & Hatamlou, A. (2016). Multi-verse optimizer: a nature-inspired algorithm for global optimization. *Neural Computing and Applications*, 27, 495-513.
- [13] Biswas, S., Kaiser, M. S., & Mamun, S. A. (2015, May). Applying ant colony optimization in software testing to generate prioritized optimal path and test data. In *2015 International Conference on Electrical Engineering and Information Communication Technology (ICEEICT)* (pp. 1-6). IEEE.
- [14] Kun, W., & Yichen, W. (2016, January). Software test case generation based on the fault propagation path coverage. In *2016 Annual Reliability and Maintainability Symposium (RAMS)* (pp. 1-4). IEEE.
- [15] Jain, M., & Gopalani, D. (2016, February). Aspect oriented programming and types of software testing. In *2016 Second International Conference on Computational Intelligence & Communication Technology (CICT)* (pp. 64-69). IEEE.
- [16] Ryalat, M. H. (2022, January). A New Algorithm to Find The K th Smallest Element in an Unordered List (Efficient for Big Data). In *2022 2nd International Conference on Computing and Information Technology (ICCIIT)* (pp. 51-56). IEEE.
- [17] Fakhouri, S. N., Hudaib, A., & Fakhouri, H. N. (2020). Enhanced optimizer algorithm and its application to software testing. *Journal of Experimental & Theoretical Artificial Intelligence*, 32(6), 885-907.
- [18] Hamad, F., Fakhuri, H., & Abdel Jabbar, S. (2022). Big data opportunities and challenges for analytics strategies in Jordanian Academic Libraries. *New Review of Academic Librarianship*, 28(1), 37-60.
- [19] Martínez-Fernández, S., Bogner, J., Franch, X., Oriol, M., Siebert, J., Trendowicz, A., ... & Wagner, S. (2022). Software engineering for AI-based systems: a survey. *ACM Transactions on Software Engineering and Methodology (TOSEM)*, 31(2), 1-59.
- [20] Li, Z., Li, T., Wu, Y., Yang, L., Miao, H., & Wang, D. (2021). Software defect prediction based on hybrid swarm intelligence and deep learning. *Computational Intelligence and Neuroscience*, 2021.
- [21] Dorgham, O., Naser, M. A., Ryalat, M. H., Hyari, A., Al-Najdawi, N., & Mirjalili, S. (2022). U-NetCTS: U-Net deep neural network for fully automatic segmentation of 3D CT DICOM volume. *Smart Health*, 26, 100304.
- [22] Rosales Muñoz, A. A., Grisales-Noreña, L. F., Montano, J., Montoya, O. D., & Perea-Moreno, A. J. (2022). Application of the Multiverse Optimization Method to Solve the Optimal Power Flow Problem in Alternating Current Networks. *Electronics*, 11(8), 1287.
- [23] Pandya, S., & Jariwala, H. R. (2022). Single-and multiobjective optimal power flow with stochastic wind and solar power plants using moth flame optimization algorithm. *Smart Science*, 10(2), 77-117.
- [24] Abualigah, L. (2020). Multi-verse optimizer algorithm: a comprehensive survey of its results, variants, and applications. *Neural Computing and Applications*, 32(16), 12381-12401.
- [25] Pachouly, J., Ahirrao, S., Kotecha, K., Selvachandran, G., & Abraham, A. (2022). A systematic literature review on software defect prediction using artificial intelligence: Datasets, Data Validation Methods, Approaches, and Tools. *Engineering Applications of Artificial Intelligence*, 111, 104773.
- [26] Hejderup, J., & Gousios, G. (2022). Can we trust tests to automate dependency updates? a case study of java projects. *Journal of Systems and Software*, 183, 111097.
- [27] Zheng, W., Shen, T., Chen, X., & Deng, P. (2022). Interpretability application of the Just-in-Time software defect prediction model. *Journal of Systems and Software*, 188, 111245.
- [28] Rath, S. K., Sahu, M., Das, S. P., & Pradhan, J. (2022). Survey on Machine Learning Techniques for Software Reliability Accuracy Prediction. In *Meta Heuristic Techniques in Software Engineering and Its Applications: METASOFT 2022* (pp. 43-55). Cham: Springer International Publishing.
- [29] Rubert, M., & Farias, K. (2022). On the effects of continuous delivery on code quality: A case study in industry. *Computer Standards & Interfaces*, 81, 103588.
- [30] Battina, D. S. (2019). Artificial intelligence in software test automation: a systematic literature review. *International Journal of Emerging Technologies and Innovative Research (www.jetir.org| UGC and issn Approved)*, ISSN, 2349-5162.
- [31] Rahkar Farshi, T., & Orujpour, M. (2021). A multi-modal bacterial foraging optimization algorithm. *Journal of Ambient Intelligence and Humanized Computing*, 1-15.
- [32] Ahmed, R., Mahadzir, S., & Mohammad Rozali, N. E. (2022, November). A Meta Model Based Particle Swarm Optimization for Enhanced Global Search. In *International Conference on Artificial Intelligence for Smart Community: AISC 2020*, 17-18 December, Universiti Teknologi Petronas, Malaysia (pp. 935-944). Singapore: Springer Nature Singapore.
- [33] Marco, R., Ahmad, S. S. S., & Ahmad, S. (2022). Bayesian hyperparameter optimization and Ensemble Learning for Machine Learning Models on software effort estimation. *International Journal of Advanced Computer Science and Applications*, 13(3).

# EFASFMM: A Unique Approach for Early Prediction of Type II Diabetics using Fire Fly and Semi-supervised Min-Max Algorithm

B. Manikyala Rao<sup>1</sup>, Mohammed Ali Hussain<sup>2</sup>

Research Scholar, Dept. of CSE, Koneru Lakshmaiah Education Foundation,  
Vaddeswaram Guntur Dist., A.P., India, Pincode: 522 302<sup>1</sup>

Professor, Dept. of Electronics and Computer Engineering,  
Koneru Lakshmaiah Education Foundation, Vaddeswaram Guntur Dist., A.P., India, Pincode: 522 302<sup>2</sup>

**Abstract**—Non-insulin-reliant, one of the most serious illnesses is diabetes mellitus, often known as type 2 diabetes, and it affects a large number of people. Between 2 and 5 million individuals worldwide die from diabetes each year. If diabetes is identified sooner, it can be managed, and catastrophic dangers including nephropathy, heart stroke, and other conditions linked to it can be avoided. Therefore, early diabetes diagnosis aids in preserving excellent health. Machine learning (ML), which has recently made strides, is now being used in a number of medical health-related fields. The innovative, nature-inspired Firefly algorithm has been shown to be effective at solving a range of numerical optimization issues. While using alliterations, the traditional firefly method employed a fixed step size models for semi-supervised learning (SSL). The firefly is effective for solving classification issues involving both a sizable number of unlabelled data and a limited number of samples with labels. The fuzzy min-max (FMM) family of neural networks in this regard provide the capability of online learning for tackling both supervised and unsupervised situations. Using a special mix of the two proposed algorithms, one of which is utilised for optimization and the other for making early predictions of type 2 diabetes. The findings for the training and testing phases for the parameter's accuracy, precision, sensitivity, specificity, and F-score are reported as 97.96%, 97.82%, 98.10%, 97.82%, and 97.95% which, when compared to current state-of-the-art methods, are finer.

**Keywords**—Fire Fly Algorithm (FFA); machine learning (ML); Semi-supervised Min-Max (SSMM)

## I. INTRODUCTION

Diabetes mellitus (DM) is the collective metabolic disorder in which people have high blood sugar levels, either because their pancreas is incapable of producing enough insulin or because their cells are unable to react to the insulin that is generated. This leads to a number of medical conditions such as polydipsia, polyuria, and polyphagia. DM is still a problem for public health everywhere in the world. This is increasingly the leading cause of death in affluent nations and is now ranked fourth or fifth among non-communicable diseases globally. In the entire world, 300 million people are predicted to have diabetes or be at risk for developing it by 2025. In the past few years, developing nations like India have had the highest growth in DM. There were 425 million diabetics worldwide as of 2017 [1], and research by the International Diabetes Federation in 2017 [2] predicted that number will rise to 625

million by 2045. Diabetes mellitus is a collection of endocrine disorders characterized by decreased glucose absorption and brought on by absolute or relative insulin deficiency. In addition to a chronic history, the disease is characterized by a disruption of all forms of metabolism. When our body's blood sugar, also known as blood glucose, is too high, we have diabetes mellitus (DM). People can develop diabetes at any age, and there are three main types: type 1, type 2, and gestational diabetes. Many hormonal and other changes that take place in the body during pregnancy are thought to be the cause of gestational diabetes, whereas other women see an increase in insulin resistance.

Type II diabetes is a chronic metabolic disorder that affects millions of people worldwide. It occurs when the body either does not produce enough insulin or is unable to use the insulin it produces effectively, resulting in high levels of glucose in the blood. Early detection of type II diabetes is crucial for effective management and prevention of complications. Predictive modeling techniques have shown great promise in identifying individuals who are at a high risk of developing type II diabetes before clinical symptoms appear. By leveraging machine learning algorithms and advanced analytics, healthcare professionals can identify individuals who are at risk of developing type II diabetes and implement preventative measures to reduce the likelihood of disease progression. This can ultimately lead to improved health outcomes and a reduction in healthcare costs associated with managing type II diabetes.

The entire paper is divided into five sections where Section I consists of introduction. Section II deals with related works, Section III deals with experimental results, Section IV deals with results of the work, and Section V deals with conclusion of the work.

## II. RELATED WORKS

Mohebbi et al. demonstrated that it was possible to use CGM signals to detect T2D patients in [3], where they offered a unique deep learning approach for the identification of type 2 diabetes. To solve the difficulties of implementing DL approaches. The authors of [4] concentrated their talks on generalised approaches, reinforcement learning, natural

language processing, deep learning in computer vision, and healthcare today (Table I).

TABLE I. VIEWS OF VARIOUS AUTHORS

Author	Contribution	Methodology	Advantages	Limitations
DPML [5]	Prediction of Type 2 diabetics	SVM, XGBoost, RF, LR	Early prediction	Works for only specific datasets not for all existing datasets
FFCSA [6]	Classification of diabetics	KNN Classifier	Notable accuracy of considered parameters	Works for static datasets not for existing
ADNNC [7]	random test and trail	DNN	High accuracy	Computation time
EDDN [8]	Prediction and classification	DNN	High accuracy	Works for limited size of data

**EFA: Firefly algorithm**

```

Begin
  Objective function  $f(x), x = (x_1, x_2, \dots, \dots, x_d)^T$ 
  Light intensity  $f_p = f(x_p)$  i.e.,  $I(r) = I_0 e^{-\gamma r^2}$ 
  Light absorption coefficient  $\gamma$ 
  While ( $t < \text{Max generation}$ ) do
    For  $s x_p \in P$  do
      For each  $x_q \in q$  do
        If  $(x_p) < f(x_q)$ , firefly  $p$  towards  $q$ 
        Vary  $\beta$  with distance  $r$  via  $\exp(-\gamma r_{pq}) \|X_p - X_q\| = \sqrt{\sum_{k=1}^d (x_{p,k} - x_{q,k})^2}$ 
         $X_{p,t+1} = X_{p,t} + \beta_0 \cdot e^{-\gamma r_{pq}^2} \cdot (X_{q,t} - X_{p,t})$ 
         $X_p = X_p + \beta_0 \cdot e^{-\gamma r_{pq}^2} \cdot (X_q - X_p) + \alpha (rand - \frac{1}{2})$ 
        Evaluate the solution and update the light intensity
      end
    end
  Fireflies are ranked and current global best  $x_p^{max}$ 
end
end
  
```

In population-based optimization techniques, efficient global exploration and local exploitation control is essential for locating the ideal solution. Therefore, during the initial stages of optimization, it is preferable to encourage individuals to roam around the whole search space rather than grouping around local optima. In order to quickly locate the best solution, it is crucial to increase convergence toward the global optima during the last stages.

**EFA: Firefly algorithm**

**Input :** UCI diabetic dataset

**Output :** Type 2 diabetic classification

**For each considered sample do**

Hyperbox membership is computed

$$b_j(A_h, V_q, W_q) = \frac{1}{2d} \sum_{p=1}^d [\max(0, 1 - \max(0, \gamma \min(1, a_{hp} - w_{qp}))) + \max(0, 1 - \max(\gamma \min(1, V_{pq} - a_{hp}), 0))]$$

For the criteria to be satisfied

$$\sum_{p=1}^d (\max(w_{q^*p}, a_{hp}) - \min(V_{q^*p}, a_{hp})) \leq d\theta$$

All hyperboxes belonging to other classes are represented as

If  $v_{q^*p} < v_{qp} \leq w_{q^*p} < w_{qp}$ , then  $\delta^{new} = \min(\min(w_{rp} - v_{q^*p}, w_{q^*p} - v_{q^*p}), \delta^{old})$

else

if

$v_{qp} < v_{q^*p} < w_{q^*p} < w_{qp}$ , then  $\delta^{new} = \min(w_{q^*p} - V_{qp}, \delta^{old})$

else

if

$v_{q^*p} < v_{kp} \leq w_{q^*p} < w_{kp}$ , then  $\delta^{new} = \min(\min(w_{q^*p} - v_{rp}, w_{kp} - v_{q^*p}), \delta)$

Overlap exists then contraction begins

If  $V_{q^*p} < V_{rp}$  then,  $v_{q^*p} < V_{rp} < w_{q^*p}$ ,  $\delta^{new}$

If  $V_{qp} < V_{q^*p} < W_{q^*p} < W_{rp}$  then  $w_{q^*p}^{new} = v_{rp}^{new} = \frac{v_{rp}^{old} + w_{q^*p}^{old}}{2}$

else

If  $v_{qp} < v_{q^*p} < w_{q^*p} < w_{qp}$ , then  $\delta^{new} = \min(w_{q^*p} - V_{qp}, \delta^{old})$

else

If  $v_{q^*p} < v_{kp} \leq w_{q^*p} < w_{kp}$ , then  $\delta^{new} =$

$\min(\min(w_{q^*p} - v_{rp}, w_{kp} - v_{q^*p}), \delta^{old})$

else

If  $V_{rp} < V_{q^*p} < W_{rp} < W_{q^*p}$  then,  $w_{rp}^{new} = v_{rp}^{new} = \frac{w_{rp}^{old} + v_{q^*p}^{old}}{2}$

Current sample encoding using new hyperbox,  $V_{q+1} =$

$W_{q+1} = A_h$

End

The algorithm provided appears to be a variant of the Fuzzy ARTMAP algorithm, which is a type of artificial neural network that uses fuzzy logic to perform pattern recognition and classification. The algorithm takes as input the UCI diabetic dataset and aims to perform classification to predict whether an individual is likely to have type II diabetes. For each sample, the algorithm computes hyper box membership using a formula that takes into account the attributes of the sample and the weights associated with the hyper boxes. The

hyper boxes represent regions in the input space that are assigned to a particular class (in this case, type II diabetes or not). The criteria for a hyper box to be considered a member of a class is that it must satisfy a certain threshold (represented by the value  $\Theta$  in the algorithm). If a hyper box overlaps with a hyper box belonging to a different class, the algorithm performs a contraction step to adjust the boundaries of the hyper boxes to reduce overlap. If the hyper box is not a member of any class, the algorithm creates a new hyper box to represent the current sample. This new hyper box is initialized with the attributes of the current sample, and the classification decision is based on whether the new hyper box satisfies the classification criteria. Overall, the algorithm is a type of supervised machine learning algorithm that uses fuzzy logic and neural networks to perform classification of the input data. It is designed to be able to adapt to new input data and adjust the hyper boxes accordingly.

Over the past ten years, machine learning and data categorization have paid a lot of attention to semi-supervised learning (SSL). Finding the target class (i.e., label) to which a data sample belongs is referred to as data classification. To do this, a collection of labelled data samples are used for model training, and the input samples are then mapped to the associated classes using the underlying learning technique. The new, unseen test samples are then categorized into the appropriate classes using the trained model. A hyper box membership function is defined as

$$f(x, \gamma) = \begin{cases} 1, & \text{if } x\gamma > 1 \\ x\gamma & \text{if } 0 \leq x\gamma \leq 1 \\ 0 & \text{if } x\gamma < 0 \end{cases}$$

Each supervised and unsupervised FMM network has a different topology made up of a variety of hyper boxes that were built up gradually (see Fig. 1). Each hyper box creates a feature boundary in a d-dimensional unit cube and is represented by a set of minimum and maximum vertices ( $I^d$ ). The value of  $[0, 1]$  determines the hyper box size; a tiny result in several hyper boxes, each of which has a small size, and vice versa.

Early prediction of Type 2 Diabetes is a growing area of research and many studies have been conducted to explore this topic. One unique approach for early prediction of Type 2 Diabetes is the use of artificial intelligence and machine learning algorithms. For example, a recent study by Wang et al. [9] (2021) used machine learning algorithms to predict the risk of Type 2 Diabetes in a Chinese population based on demographic and lifestyle factors. Another study by Noh et al. (2021) [10] used artificial intelligence to predict the risk of Type 2 Diabetes based on clinical data from electronic health records.

In addition, several studies have explored the use of biomarkers for early prediction of Type 2 Diabetes. For instance, a study by Senn et al. (2020) [11] investigated the use of a blood-based biomarker called plasma branched-chain amino acids ( BCAAs) to predict the risk of Type 2 Diabetes in a cohort of Finnish individuals. Similarly, another study by Li et al. (2020) [12] explored the use of urinary metabolites as a biomarker for early prediction of Type 2 Diabetes in a Chinese

population. Furthermore, there has been research into the use of genetic information for early prediction of Type 2 Diabetes. A study by Wang et al. (2020) [13] investigated the use of genetic risk scores to predict the risk of Type 2 Diabetes in a Korean population. Another study by Bancks et al. (2021) [14] used a genetic risk score and lifestyle factors to predict the risk of Type 2 Diabetes in a diverse population in the United States. Overall, these studies demonstrate the potential of various approaches for early prediction of Type 2 Diabetes, including the use of artificial intelligence and machine learning algorithms, biomarkers, and genetic information.

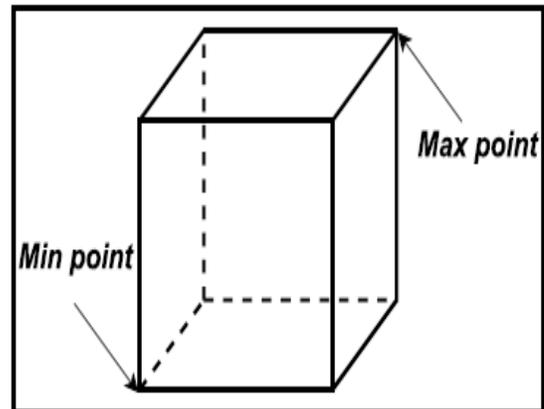
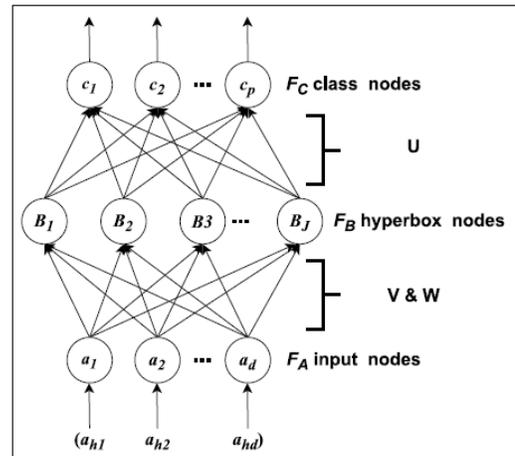


Fig. 1. (a), (b) Supervised FMM network and a 3D hyperbox.

### III. EXPERIMENTAL SET UP

Utilizing the Visual Studio IDE, a specific User Interface (UI) is created in order to connect to the server and obtain performance metrics for both existing and suggested methods. The proposed system is implemented in Visual C++ and the existing methods are implemented in Common Language Runtime (CLR) libraries. The UI is operated on a machine that connects to the server using an 8 GB RAM and an Intel Core i5-7200 processor operating at 2.7 GHz. Through the I2K2 cloud server intermediary infrastructure, Amazon Server is rented. The server has 100GB of High IOPS Solid State Hard Drive (SSD), 2GB of RAM, 2 Virtual CPU Computational Cores, and a 99.99% uptime guarantee. For Single Admin Windows Operating System and Single User Remote Desktop Server (RSD) Client Access License (CAL) included, the

software licenses are provided by the I2K2 service. Dataset [15] is considered to carry out the results (see Fig. 2).

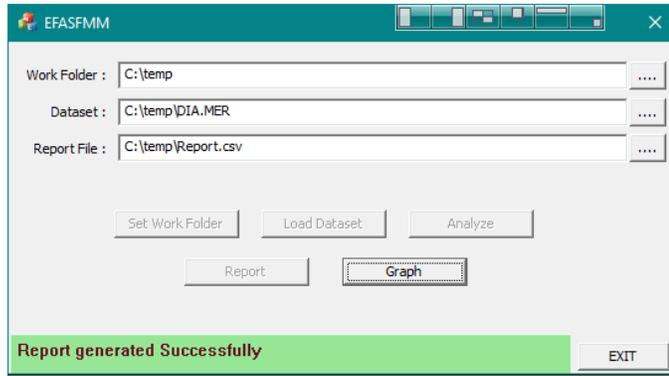


Fig. 2. UI for running the proposed approach.

#### IV. RESULTS

The model is trained and tested with the dataset with a ratio of 60-40. The results are considered in terms of accuracy, precision, specificity, sensitivity and F-Score.

##### A. Accuracy

TABLE II. ACCURACY OBTAINED DURING TRAINING

Data (%)	DPMLA	FFCSA	DNNC	EDDN	EFASFMM
7	38.54	33.40	36.54	43.72	45
14	53.89	47.28	54.93	59.69	61.13
21	63.10	55.69	65.63	69.43	70.19
28	69.38	61.38	73.37	76.02	76.99
35	74.51	66.30	79.06	81.39	82.34
42	78.64	69.66	83.98	85.49	86.64
49	81.87	72.99	87.99	88.71	90.01
56	84.60	75.49	91.50	92.04	93.17
63	87.53	78.12	94.27	94.96	96.15
70	89.90	80.31	97.27	97.27	98.75

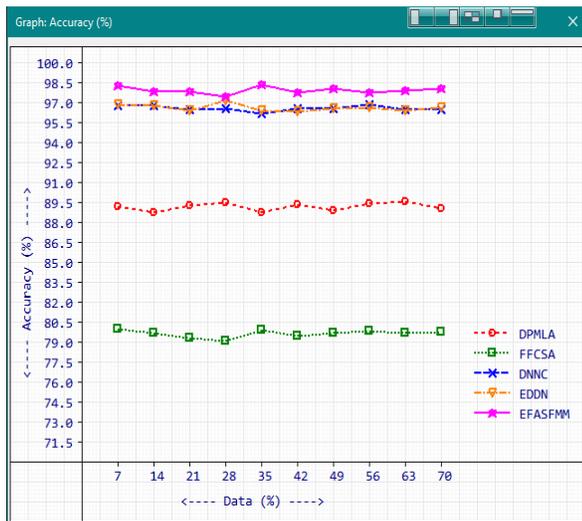


Fig. 3. Graphs of accuracy obtained during training.

TABLE III. ACCURACY OBTAINED DURING TESTING

Data (%)	DPMLA	FFCSA	DNNC	EDDN	EFASFMM
7	89.23	79.97	96.83	96.86	98.35
14	88.79	79.73	96.79	96.8	97.84
21	89.3	79.34	96.54	96.45	97.88
28	89.53	79.07	96.55	97.19	97.49
35	88.75	79.93	96.23	96.44	98.37
42	89.41	79.45	96.59	96.33	97.77
49	88.94	79.72	96.55	96.55	98.07
56	89.46	79.88	96.89	96.63	97.8
63	89.61	79.70	96.53	96.45	97.94
70	89.08	79.8	96.48	96.63	98.09

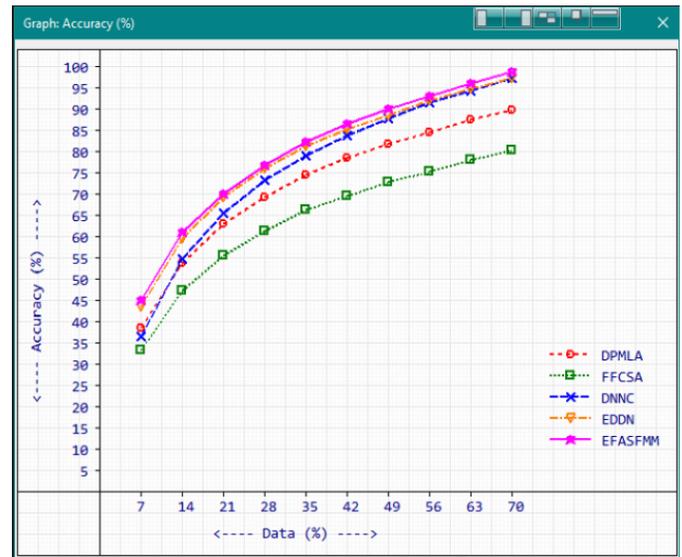


Fig. 4. Graphs of accuracy obtained during testing.

Accuracy values are referred in the Table II for training and Table III for testing. During training the average values of accuracy for the existing and proposed approaches such as DPMLA, FFCSA, DNNC, EDDN and EFASFMM observed are 72.20%, 64.06%, 76.45%, 78.87% and 80.04%. During testing the average values of accuracy for the existing and proposed approaches are given as 89.21%, 79.66%, 96.60%, 96.63% and 97.96%. Accuracy graphs are shown in Fig. 3 and Fig. 4 during training and testing.

##### B. Precision

Precision values are referred in the Table IV for training and Table V for testing. The corresponding graphs for precision during training and testing are shown in Fig. 5 and 6, correspondingly.

During training the average values of Precision for the existing and proposed approaches such as DPMLA, FFCSA, DNNC, EDDN and EFASFMM observed are 73.73%, 63.75%, 75.59%, 79.03% and 79.72%. During testing the average values of precision for the existing and proposed approaches are given as 90.68%, 79.63%, 95.45%, 96.74% and 97.82%.

TABLE IV. PRECISION OBTAINED DURING TRAINING

Data (%)	DPMLA	FFCSA	DNNC	EDDN	EFASFMM
7	40	32.45	36	44	44.45
14	55.35	46.72	54.24	60.13	60.71
21	64.87	55.17	65.17	69.65	69.76
28	70.88	60.98	72.48	76.09	76.51
35	76.01	66.00	78.30	81.59	82.28
42	80.04	69.35	82.96	85.42	86.47
49	83.37	73.01	86.88	88.79	89.81
56	86.05	75.43	90.63	92.04	92.94
63	89.21	77.98	93.25	95.12	95.80
70	91.54	80.36	96	97.54	98.45

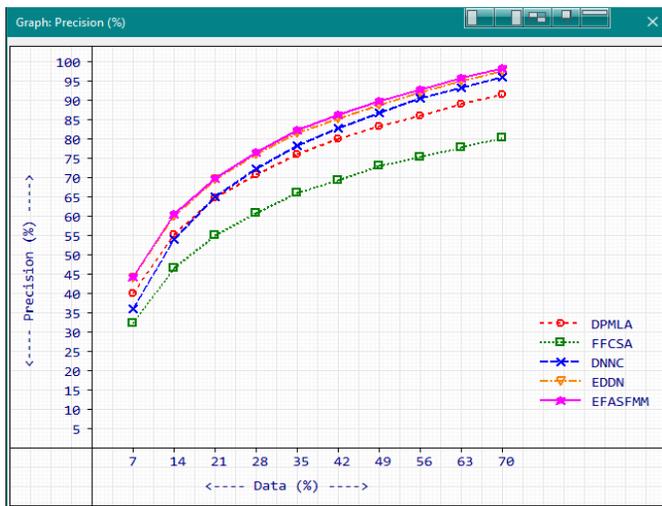


Fig. 5. Precision graph obtained during training.

TABLE V. PRECISION OBTAINED DURING TESTING

Data (%)	DPMLA	FFCSA	DNNC	EDDN	EFASFMM
7	90.45	80.04	95.39	96.58	98.17
14	90.01	79.21	95.85	96.97	98.17
21	91.11	79.38	95.71	96.67	98.22
28	90.89	79.1	95.57	97.22	97.43
35	90.49	80.01	95.12	96.52	98.04
42	91.05	79.5	95.04	96.66	97.12
49	90.42	79.45	95.51	97.08	98.11
56	90.72	80.12	95.53	96.08	97.48
63	91.18	79.92	95.3	96.75	97.49
70	90.54	79.58	95.49	96.94	97.98

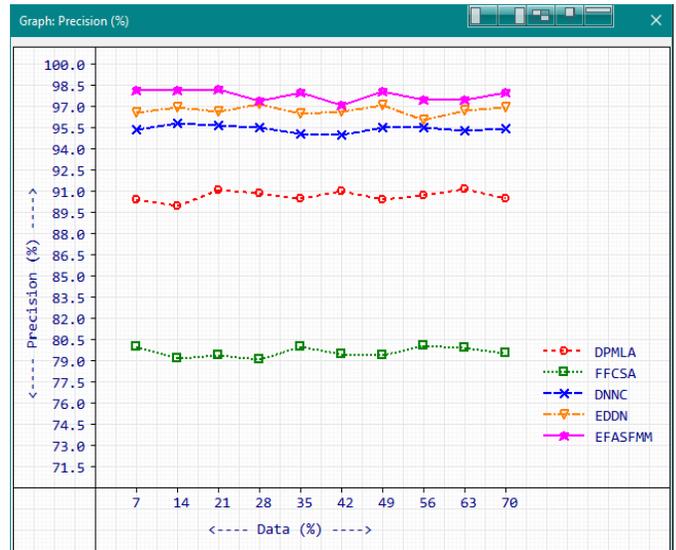


Fig. 6. Precision graph obtained during testing.

### C. Sensitivity

Sensitivity values are referred in the Table VI for training and Table VII for testing. The corresponding graphs for precision during training and testing are shown in Fig. 7 and 8, correspondingly.

During training the average values of sensitivity for the existing and proposed approaches such as DPMLA, FFCSA, DNNC, EDDN and EFASFMM observed are 71.53%, 64.07%, 76.99%, 78.80% and 80.20%. During testing the average values of sensitivity for the existing and proposed approaches are given as 88.09%, 79.68%, 97.69%, 96.53% and 98.10%.

TABLE VI. SENSITIVITY OBTAINED DURING TRAINING

Data (%)	DPMLA	FFCSA	DNNC	EDDN	EFASFMM
7	38.86	33.08	36.39	43.76	44.94
14	53.78	47.25	55.00	59.61	61.23
21	62.65	55.75	65.78	69.35	70.36
28	68.82	61.47	73.80	75.99	77.26
35	73.79	66.40	79.50	81.27	82.37
42	77.85	69.78	84.69	85.54	86.76
49	80.94	72.97	88.85	88.65	90.17
56	83.62	75.51	92.22	92.04	93.37
63	86.30	78.19	95.20	94.81	96.48
70	88.64	80.29	98.50	97.01	99.03

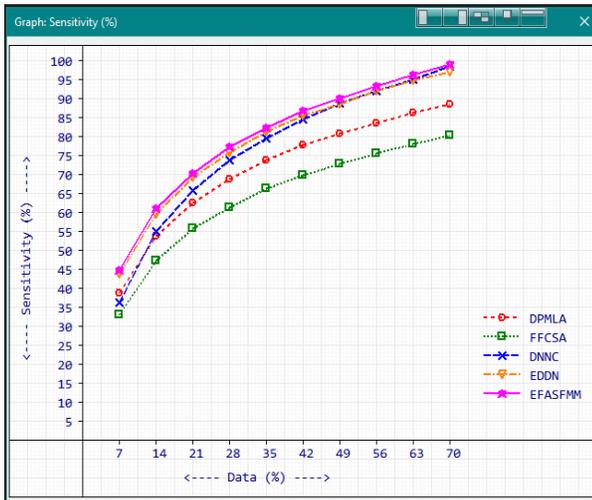


Fig. 7. Sensitivity graph obtained during training.

TABLE VII. SENSITIVITY OBTAINED DURING TESTING

Data (%)	DPMLA	FFCSA	DNNC	EDDN	EFASFMM
7	88.30	79.93	98.21	97.12	98.52
14	87.87	80.04	97.69	96.64	97.53
21	87.92	79.32	97.33	96.24	97.56
28	88.48	79.06	97.49	97.17	97.55
35	87.44	79.88	97.27	96.36	98.69
42	88.16	79.42	98.08	96.03	98.39
49	87.82	79.88	97.53	96.06	98.04
56	88.49	79.74	98.21	97.15	98.10
63	88.40	79.57	97.71	96.18	98.37
70	87.97	79.93	97.41	96.34	98.205

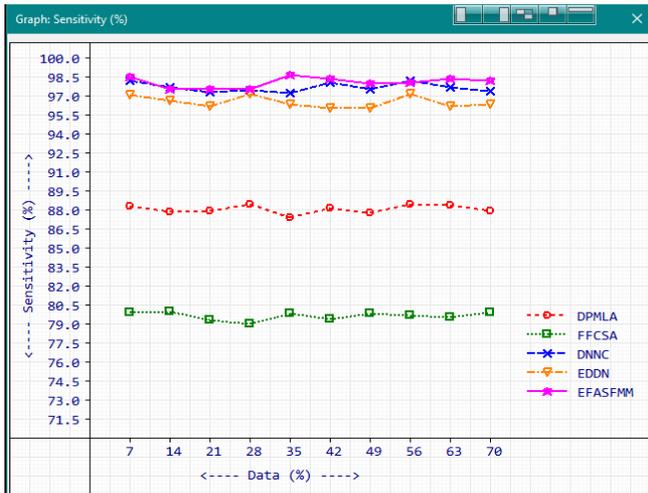


Fig. 8. Sensitivity graph obtained during testing.

#### D. Specificity

Specificity values are referred in the Table VIII for training and Table IX for testing. The corresponding graphs for precision during training and testing are shown in Fig. 9 and 10 correspondingly.

TABLE VIII. SPECIFICITY OBTAINED DURING TRAINING

Data (%)	DPMLA	FFCSA	DNNC	EDDN	EFASFMM
7	38.20	33.71	36.69	43.69	45.05
14	54.01	47.31	54.87	59.78	61.04
21	63.58	55.63	65.49	69.51	70.02
28	69.98	61.29	72.96	76.06	76.73
35	75.26	66.20	78.62	81.51	82.30
42	79.47	69.54	83.30	85.44	86.52
49	82.85	73.00	87.17	88.77	89.85
56	85.63	75.46	90.79	92.04	92.98
63	88.83	78.04	93.38	95.10	95.83
70	91.25	80.34	96.09	97.53	98.46

During training the average values of specificity for the existing and proposed approaches such as DPMLA, FFCSA, DNNC, EDDN and EFASFMM observed are 72.91%, 64.05%, 75.94%, 78.94% and 79.88%. During testing the average values of specificity for the existing and proposed approaches are given as 90.40%, 79.64%, 95.55%, 96.74% and 97.82%.

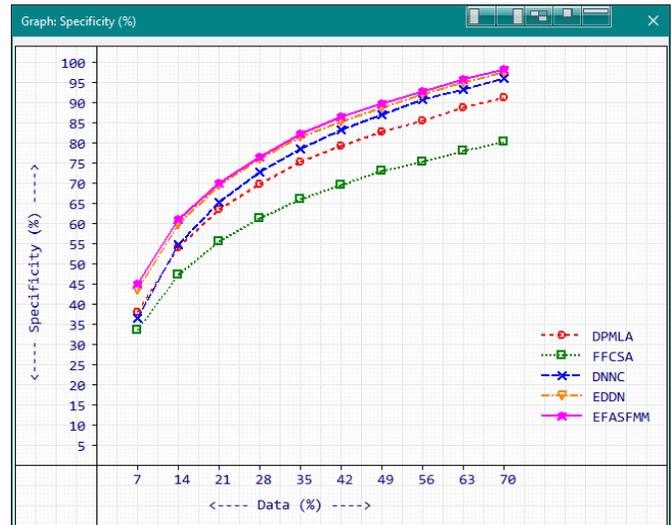


Fig. 9. Specificity graph obtained during training.

TABLE IX. SPECIFICITY OBTAINED DURING TESTING

Data (%)	DPMLA	FFCSA	DNNC	EDDN	EFASFMM
7	90.21	80.01	95.51	96.59	98.17
14	89.76	79.42	95.92	96.95	98.15
21	90.77	79.36	95.78	96.65	98.20
28	90.63	79.08	95.65	97.21	97.43
35	90.14	79.97	95.22	96.51	98.05
42	90.74	79.48	95.18	96.63	97.15
49	90.12	79.56	95.60	97.04	98.10
56	90.48	80.02	95.64	96.12	97.49
63	90.89	79.83	95.41	96.73	97.51
70	90.25	79.66	95.57	96.92	97.98

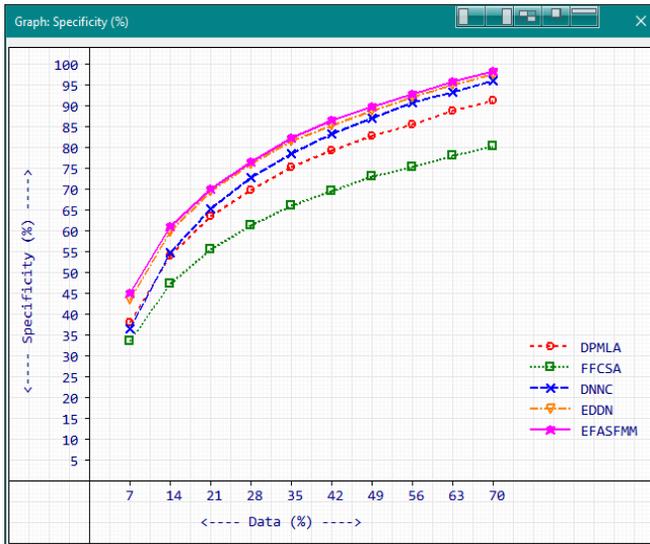


Fig. 10. Specificity graph obtained during testing.

E. F-Score

F-Score values are referred in the Table X for training and Table XI for testing. The corresponding graphs for precision during training and testing are shown in Fig. 11 and 12 correspondingly.

TABLE X. F-SCORE OBTAINED DURING TRAINING

Data (%)	DPMLA	FFCSA	DNNC	EDDN	EFASFMM
7	39.42	32.76	36.19	43.88	44.69
14	54.55	46.98	54.62	59.87	60.96
21	63.74	55.46	65.47	69.50	70.06
28	69.83	61.23	73.13	76.04	76.88
35	74.88	66.20	78.90	81.43	82.33
42	78.93	69.56	83.82	85.48	86.62
49	82.14	72.99	87.85	88.72	89.99
56	84.82	75.47	91.42	92.04	93.16
63	87.73	78.09	94.21	94.96	96.14
70	90.07	80.32	97.23	97.28	98.74

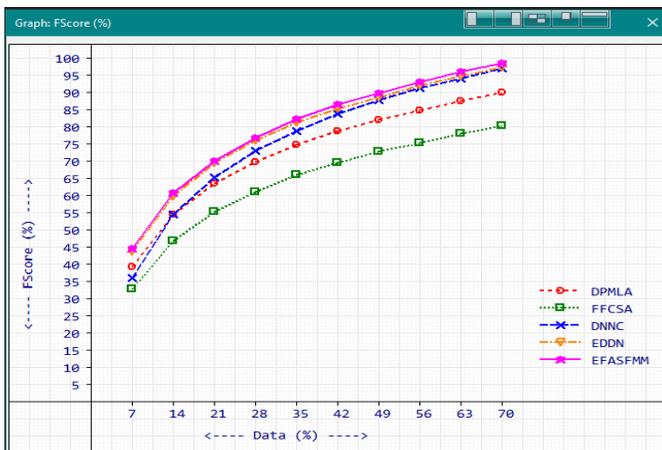


Fig. 11. F-score graph obtained during training.

TABLE XI. F-SCORE OBTAINED DURING TESTING

Data (%)	DPMLA	FFCSA	DNNC	EDDN	EFASFMM
7	89.36	79.98	96.78	96.85	98.34
14	88.92	79.62	96.76	96.80	97.85
21	89.49	79.35	96.51	96.45	97.89
28	89.67	79.08	96.52	97.19	97.49
35	88.94	79.94	96.18	96.44	98.36
42	89.58	79.46	96.53	96.34	97.75
49	89.10	79.66	96.51	96.56	98.07
56	89.59	79.93	96.85	96.61	97.79
63	89.77	79.74	96.49	96.46	97.93
70	89.23	79.75	96.44	96.64	98.09

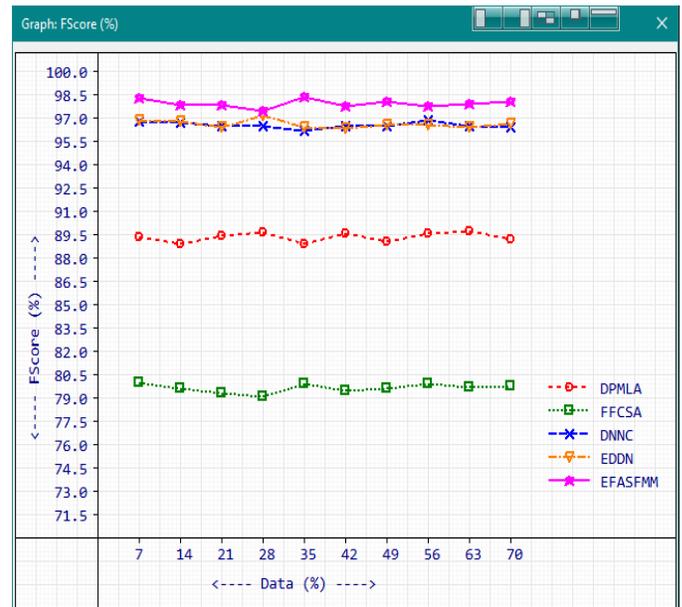


Fig. 12. F-score graph obtained during testing.

During training the average values of specificity for the existing and proposed approaches such as DPMLA, FFCSA, DNNC, EDDN and EFASFMM observed are 72.61%, 63.91%, 76.29%, 78.92% and 79.96%. During testing the average values of specificity for the existing and proposed approaches are given as 89.36%, 79.65%, 96.56%, 96.63% and 97.95%.

V. CONCLUSION

Finding the best diabetes treatment primarily depends on early illness detection. One of the most common diseases in the world is diabetes. Whatever the sort of sickness, it is a frequent problem for doctors, medical professionals, and scientists to forecast the disease in its early stages. The main cause of this is a lack of awareness in developing and underdeveloped nations. A person's life can be saved by ignoring paradoxical events, diagnosing the sickness early, and taking the right medication. The upgraded firefly technique and the semi-supervised min-max approach algorithm are combined in an original way in the current paper. Where the min-max strategy is employed for the early diagnosis of type 2 diabetes and the firefly algorithm is used for optimization. The proposed approach is unique in that

it achieves the best results in terms of the metrics taken into account, including accuracy, precision, sensitivity, specificity, and F-Score. The type 2 diabetes dataset is used in the sense of a 60-40%. The results for the factors that were taken into consideration during training and testing are presented as follows: 80.00%, 79.72%, 80.20%, 79.88%, 79.96%, and 97.96%, 97.82%, 98.10%, 97.82%, 97.95%. The suggested method is beneficial for both training and testing purposes, as well as for the early diagnosis of type 2 diabetics while taking into account minimal computational time and highly accurate findings. The primary problem facing researchers in the current situation is identifying the causes of retinopathy in type 2 diabetes.

#### REFERENCES

- [1] International Diabetes Federation, IDF Diabetes Atlas, 8th edn. (2017).
- [2] G. Li, S. Peng, C. Wang, J. Niu, Y. Yuan, An energy-efficient data collection scheme using denoising autoencoder in wireless sensor networks. *Tsinghua Sci. Technol.* 24(1), 86–96 (2019).
- [3] A. Mohebbi, B. Aradottir, R. Johansen, H. Bengtsson, M. Fraccaro, M. Morup, in Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS. A deep learning approach to adherence detection for type 2 diabetics (2017), pp. 2896–2899.
- [4] A. Esteva, A. Robicquet, B. Ramsundar, V. Kuleshov, M. DePristo, K. Chou, J. Dean, A guide to deep learning in healthcare. *Nat. Med.* 25(1), 24–29 (2019).
- [5] Charitha, C., et al. "Type-II Diabetes Prediction Using Machine Learning Algorithms." 2022 International Conference on Computer Communication and Informatics (ICCCI). IEEE, 2022.
- [6] Haritha, R., D. Suresh Babu, and P. Sannulal. "A hybrid approach for prediction of type-1 and type-2 diabetes using firefly and cuckoo search algorithms." *International Journal of Applied Engineering Research* 13.2 (2018): 896-907.
- [7] Nadesh, R. K., and K. Arivuselvan. "Type 2: diabetes mellitus prediction using deep neural networks classifier." *International Journal of Cognitive Computing in Engineering* 1 (2020): 55-61.
- [8] Zhou, Huaping, Raushan Myrzashova, and Rui Zheng. "Diabetes prediction model based on an enhanced deep neural network." *EURASIP Journal on Wireless Communications and Networking* 2020.1 (2020): 1-13.
- [9] Wang Y, Cheng J, Chen L, Chen Y, Chen Y, Jiang Y, et al. Prediction of type 2 diabetes risk in Chinese population using machine learning algorithms. *BMC Endocr Disord.* 2021 Feb 17;21(1):28.
- [10] Noh J, Kim H, Lee J, Kim Y, Kim C, Kim K, et al. Deep learning-based risk prediction model for type 2 diabetes mellitus. *Sci Rep.* 2021 Jan 21;11(1):1962.
- [11] Senn T, Hazeghazam M, Viljakainen H, Kröger H, Poutanen K, Schwab U, et al. Plasma Branched-Chain Amino Acids as a Predictor of Risk of Type 2 Diabetes in Finnish Men. *Diabetes Care.* 2020 Dec;43(12):3034-3041.
- [12] Li Y, Wang T, Wei Y, Xu L, Zhao X, Wang W, et al. Urinary metabolites as a predictor of incident type 2 diabetes mellitus among Chinese adults. *Diabetologia.* 2020 Sep;63(9):1933-1943.
- [13] Wang J, Kim Y, Lee J, Kim Y, Park T. Development and validation of a Korean diabetes prediction model using a genome-wide association study. *BMC Med Genet.* 2020 Jun 12;21(1):126.
- [14] Bancks MP, Kanaya AM, Kandula NR, Chang YF, Huang ES. Development and validation of a prediction model for risk of incident type 2 diabetes in a multi-ethnic cohort. *PLoS One.* 2021 Mar 8;16(3):e0248282.
- [15] <https://archive.ics.uci.edu/ml/datasets/diabetes>.

# Development of a Mobile Application to Reduce the Rate of People with Text Neck Syndrome

Rosa Perez-Siguas, Hernan Matta-Solis, Eduardo Matta-Solis, Hernan Matta-Perez, Luis Perez-Siguas, Randall Seminario Unzueta, Victoria Tacas-Yarcuri

TIC Research Center: eHealth & eEducation, Instituto Peruano de Salud Familiar, Lima-Peru

**Abstract**—Now-a-days, it is no surprise that mobile devices have become a very useful tool in the daily tasks of many people worldwide. This is thanks to their various features such as portability, connectivity, entertainment, work tool, etc. However, due to the bad posture that users have when using them, a syndrome called "Text Neck" is produced. This is caused by prolonged use of the devices looking down and tilting the head at different angles. The degree of inclination of the head causes a detrimental effect on the neck joints, so that the greater the degree of inclination the effect of the weight of the head on the neck increases detrimentally. However, currently mobile devices have sensors that help in monitoring the activities of users, in this sense, there is the gyroscope that allows the completion of the position and the accelerometer that tells us the amount of movement of the device. In this sense, a mobile application has been developed that by monitoring the information of the angle of inclination of the device and the time it remains in the same, allows notifying users to adopt a proper position. The aim is to reduce the number of people affected by text neck syndrome.

**Keywords**—Accelerometer; android; firebase; gyroscope; mobile devices; sensors; text neck

## I. INTRODUCTION

Today, we find that mobile devices such as cell phones, tablets or e-books have become essential attachments in the daily lives of millions of people worldwide [1]–[6]. According to statistical data collected in 2021, the number of mobile device users will exceed 3.8 million, representing 48.53% of the world's population [1]. This popularity in the massive use of mobile devices is due to the various activities that can be performed with them, in this sense, users can exchange information, access the Internet, play in mobile applications, and other types of activities [5], [7], [8] They also allow for fluid communication between users through the use of text messages and social networks [1]–[3], [8]

However, beyond the popularity that mobile devices have achieved, due to prolonged use and poor posture of users, health-related problems have been identified [1]–[10]. The problem is now known as "Text Neck" syndrome, a term coined by American chiropractor Dr. Dean L. Fishman [1], [3], [6]–[8]. However, this syndrome is produced due to the downward inclination of the users' neck and the excessive forward bending when using mobile devices [1], [3]–[7], [10]. In this sense, the greater the angle of inclination, the greater the

weight that falls on the neck joints, being 12 kg at an angle of 15 degrees, 18 kg at an angle of 30 degrees, 22 kg at an angle of 45 degrees, and 27 kg at an angle of 60 degrees [1], [2], [5], [6], [9], [10]. As a result, the user may experience many ailments ranging from neck pain, neck stiffness, reduced mobility, headaches, postural disturbances, rounded shoulders [1], [6], [7], [10].

On the other hand, nowadays cellular devices have a wide range of sensors that allow the detection and monitoring of many activities [11]–[13]. In this sense, the accelerometer and the gyroscope can be found as complementary sensors in the measurement of certain activities [11], [14]–[17]. Therefore, due to the precision that cell phones possess [13], [18], activities such as swaying in the elderly can be monitored [19], establish the frequency of the footprint [14], detection of abnormal behaviors [11], determine the degree of sedentary lifestyle of the users [17] and video stabilization through the use of the gyroscope [20]. Now, these sensors can be used to detect the time that a person has been using the cell phone and the position in which it is used. This information can be provided by the accelerometer and gyroscope, then through an application developed in Android can be analyzed for the purposes of the case [11], [12], [14], [15], [19], [20].

Therefore, taking into account the information that can be obtained through the sensors of mobile devices, we developed a mobile application that allows the monitoring of the position in which it is used by users through the gyroscope; the time the user uses the device in a specific position taking into account the viewing angles. The purpose of our research is the development of a mobile application to notify users of the excessive use of mobile devices and the position in which it is used, with the aim of reducing the rate of people suffering from text neck syndrome.

Next, the activities carried out within the present research project are detailed, in this sense, in Section II we specify the methodology that has been used to obtain the main objective of our project, in Section III we show the results obtained after implementing our application, in Section IV we made a comparison of the contributions obtained with those of other authors, and finally in Section V we show the conclusions that we have reached after having carried out the present research project.

## II. METHODOLOGY

### A. Determination of Angle of Inclination

Text neck syndrome is characterized by the angle of inclination generated in the user's neck when tilting the head forward to view mobile devices [1], [5], [7], [9], [10], [21]. Thus, as shown in Table I, as the degree of inclination of the user's neck increases, there is a direct effect on the weight exerted on the spine [1].

TABLE I. HEAD WEIGHT ACCORDING TO THE ANGLE OF INCLINATION

Angle of inclination of the neck	Force exerted on the neck
0°	4,54 – 5,44 kg
15°	12,25 kg
30°	18,14 kg
45°	22,23 kg
60°	27,22 kg

Likewise, Fig. 1 shows graphically the angle of inclination of the neck forward and the weight effect equivalent to each one for the reader's better understanding.

Taking into account that the sensors are located inside the mobile devices, the detection of the information of the users' neck inclination angles must be related to the inclination angles of the devices where our device is running. Consequently, the calculation of the tilt angles taken by the mobile devices in relation to the users' neck tilt angles was performed.

In that sense, the users' neck inclination angles vary in an incremental range of 15°. Likewise, the initial angle is 0° and the maximum angle of inclination is 60°, where it is possible to identify the inclusion of four increments to get from the minimum to the maximum angle of inclination.

Tilt interval range = 15°

Now, before continuing with the calculations, it was taken into account that users seek the best position of the mobile device for viewing the various desired contents. Therefore, after performing an analysis of the way in which mobile devices are held by users, a correlation was found between the angle of inclination and the angle at which the mobile device is held.

In effect, the mobile devices when held cover different angles of a quadrant. Taking into account the aforementioned, the calculation of the range of degree intervals that the devices have in the different positions was carried out, resulting in 22.5°.

Cell interval range = 22.5

The angles of inclination that the mobile device can take can be seen more clearly in Fig. 2, taking 90° as the starting point, which are traveled in intervals of 22.5° until reaching 180°. On the other hand, Table II shows the equivalence of each of the degree intervals in both the inclination of the user's neck and the angle at which the mobile device is held.

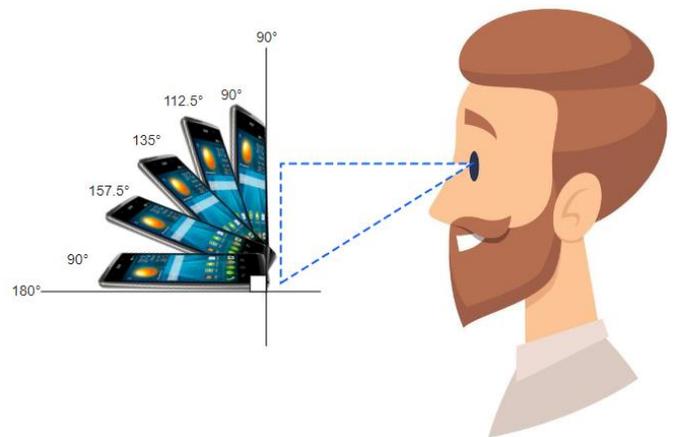


Fig. 1. Tilt angles of the mobile devices

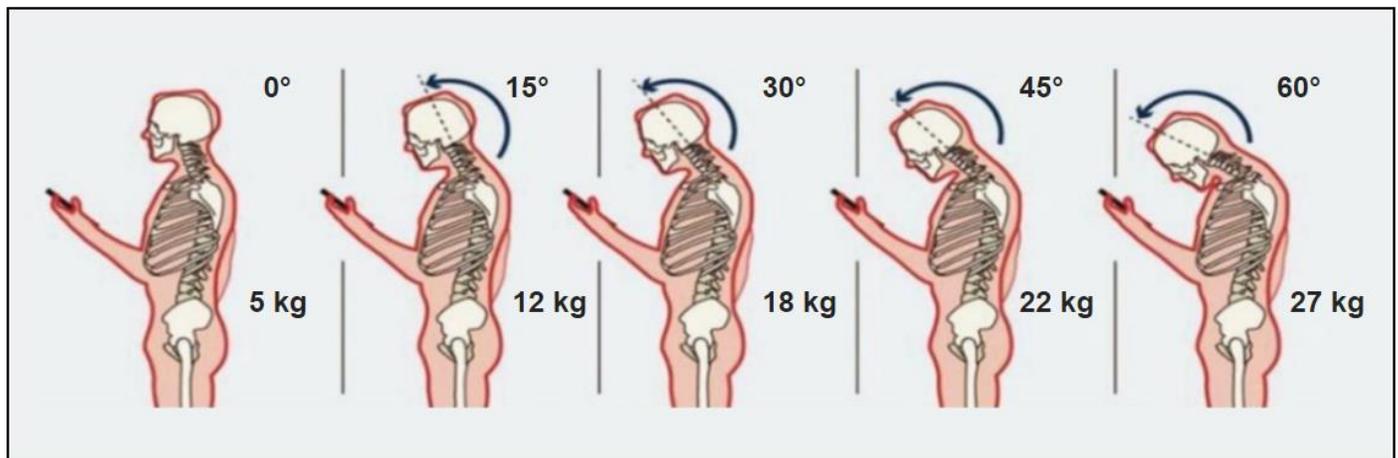


Fig. 2. Angle of inclination of the neck and the equivalent weight effect on the spine

TABLE II. EQUIVALENCE OF USER ANGLES AND MOBILE DEVICES

Angle of inclination of the neck	Tilt angle of the mobile device
$0^\circ < 15^\circ$	4,54 – 5,44 kg
$15^\circ < 30^\circ$	12,25 kg
$30^\circ < 45^\circ$	18,14 kg
$45^\circ < 60^\circ$	22,23 kg
$60^\circ$	27,22 kg

**B. Flowchart**

In this section, as can be seen in Fig. 3, the actions that are evaluated within the application are detailed, as well as the path that must be followed for the evaluated actions. This diagram gives a clearer idea of the actions that must be programmed within our application, as well as determining the evaluation criteria at the time of the application's decision-making process. In this sense, the application is started when the cell phone is unlocked by the user; then the angle of inclination of the mobile device is determined by analyzing the gyroscope information; then the amount of inertia that the device carries in the position carried by the user is determined; then the application proceeds to keep track of the time in which the device is used in an inappropriate position; and finally the device makes a notification for the user to take the necessary measures to avoid the use of the device in that position.

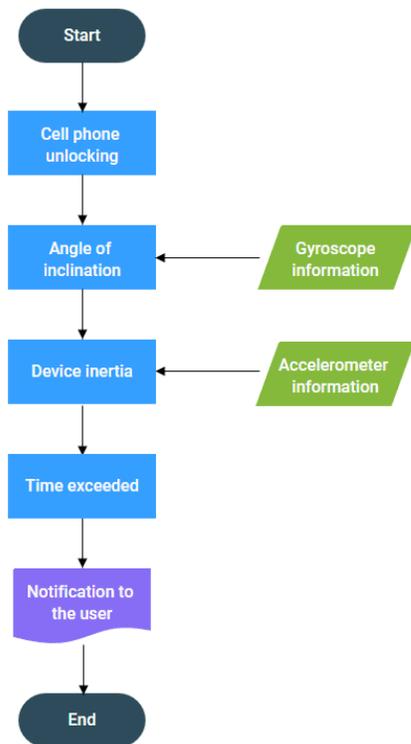


Fig. 3. System flow diagram

**C. Prototyping of the Application**

In this section, the respective prototyping of the different phases that are part of this application was carried out. In this sense, a detailed analysis of the actions that are framed in its workflow was carried out, resulting in the necessary modules

for the correct fulfillment of the objectives of this research project.

Now, our project was developed keeping a user-friendly design through the use of colors that allow the understanding of the messages. In this sense, the modules were developed:

1) *Welcome and login:* As part of the development of this module, we took into account the need to keep track of those users who make use of our application. Therefore, as can be seen in Fig. 4(A) where the Welcome module is shown, the application icon has been placed at the beginning; then a phrase that allows users to identify the purpose of the application; the login button that allows access to the login form; and finally a couple of options that ask the user about the possession of an account. Also, in Fig. 4(B) login module we include as title the name of the application; as in the previous model, the icon of the application is shown; then the data required to complete this form are shown, which are the username and password; likewise, the options to remind the user and the login button are shown; finally, the user is given the possibility to access our application through the use of their social networks such as Facebook, Google, Twitter, and LinkedIn, since they are the most used social networks today [22].

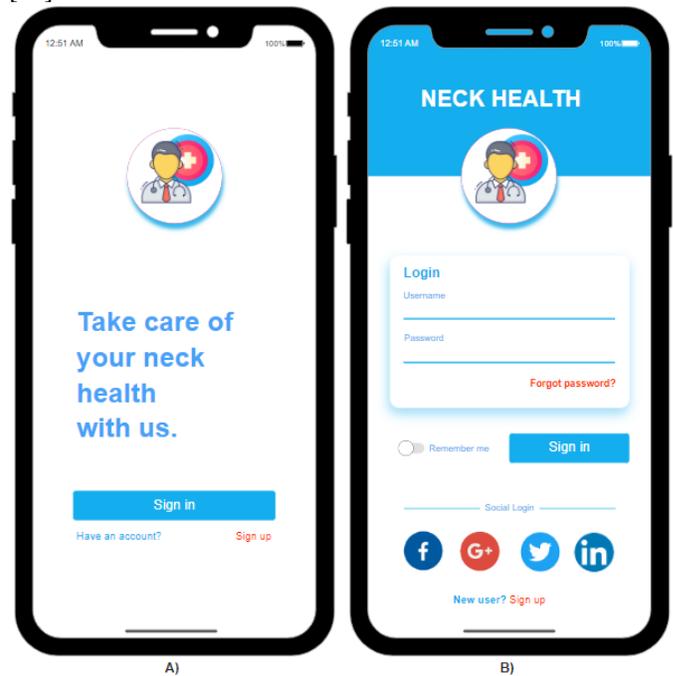


Fig. 4. A) Welcome module. B) System login module

2) *Main menu:* For the development of this module we took into account the functions to which the user requires access within our application. In this sense, as can be seen in Fig. 5 of the main menu, a couple of buttons have been implemented, the first one gives the user the possibility to access the current tilt status, and the second button allows the user to access the statistics of the use of the cell phone.

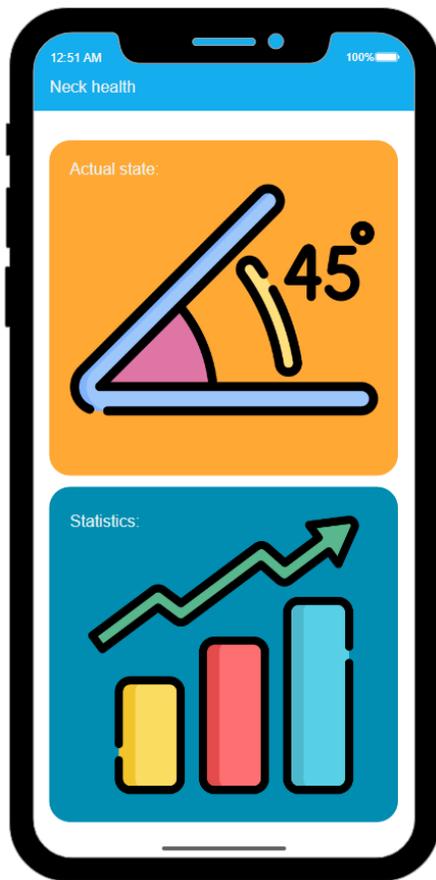


Fig. 5. Módulo de menú principal

#### D. Hardware

As part of the development of our research project, the hardware technologies necessary for the project to meet the previously stated objective were determined.

1) *Accelerometer*: This sensor allows us to obtain information on the acceleration of the device during the time it is used by the user, which makes it possible to determine the amount of inertia and other daily activities [11], [15]–[17].

2) *Gyroscope*: This sensor has the ability to detect the tilt angles in which the devices that have it implemented [15], [18], [20], this is of utmost importance since it is very much in line with the main objective of our research project.

3) *Smartphone*: Mobile devices that have a great acceptance by the general public, surpassing 80% of North American users [13]. Thus, these mobile devices have several sensors, among which are included the accelerometer and the gyroscope, both being complementary in the detection of movement and determination of activities by the users of these mobile devices [11], [12], [15], [23], [24]. Taking into account the characteristics and the utilization rate of these devices, it has been determined as the ideal device for the implementation of our research project.

#### E. Software

Within the development of this project, the following software technologies were selected due to their characteristics and features.

1) *Android*: It is an operating system that is used in a large number of mobile devices which allows you to run a large number of applications according to the needs of users [25]. It also allows the implementation of many functions for the management of user data [26]. Taking into account these characteristics, this operating system was used for the deployment of our mobile application.

2) *Android studio*: This IDE has a large number of features that allow the correct development of mobile applications for the solution of the diverse needs that users have [25]. In this sense, this IDE allowed us to carry out the correct development of the software infrastructure necessary to achieve the objective of our research project.

3) *Firebase*: This allows us to handle a large amount of unstructured data or commonly known by the name of NO-SQL [26]. This technology allows us to handle a large amount of data in real time and other features that are easy to implement in Android applications [27]–[29]. Therefore, this technology was used for the storage of user account data and usage data of their mobile devices.

### III. RESULTS

Once the development of our mobile application has been completed, we will analyze the results obtained after its implementation.

#### A. Implementation of the Application

Initially, the application was installed on a mobile device with the Android operating system. Fig. 6 shows how the application has been deployed by means of an icon that has been added to the main screen, through the use of this icon the user can access the application system.

#### B. Current status of the Device

After the user has successfully authenticated, the system will allow the user to enter the main menu of the system. At this point, the application, when used for the first time, will require the user's authorization to access certain features and data of his mobile device. Then, the user will be able to access the main menu of our application where you can see the options to which the user has access.

Now, among the options shown in the main menu is the Actual State option. This option collects the gyroscope information that is integrated inside the mobile device. Fig. 7 shows the result of the information obtained from the gyroscope sensor by means of which the angle of inclination that the device has at that moment and its equivalent angle of inclination in the user's neck can be given as a result.



Fig. 6. The application shows the angle of inclination and the equivalent weight

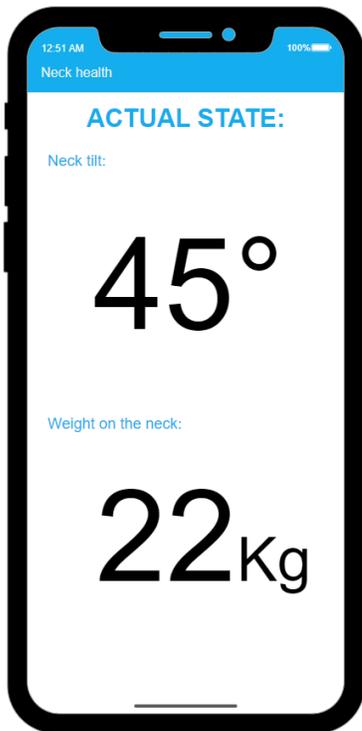


Fig. 7. Application deployment on a mobile device

This module allows the user to have immediate access to information on the tilt angles of both their device and the equivalent angle their neck is taking in relation to the tilt of the mobile device. This allows users to correct their neck position and change the way they are using their mobile devices by adapting better positions that safeguard their health.

### C. Device Usage Statistics

As part of the options implemented within the application, we find the statistics option. As shown in Fig. 8, this module presents the statistics options according to the day, week, month and year; in each of the above-mentioned cases a statistical table is presented where the frequency of incidence in the angles of inclination can be appreciated; likewise, in the lower part of the module the application shows which is the largest angle registered according to the statistical option that has been selected by the user.

This module allows users to keep track of the angles at which the mobile device is used, in this sense, in view of the data shown by the application, the user has the possibility to keep track of the change in their habits in the use of mobile devices. Therefore, the application is an extremely important tool to keep track of the angles of inclination that users have with respect to the use of mobile devices.

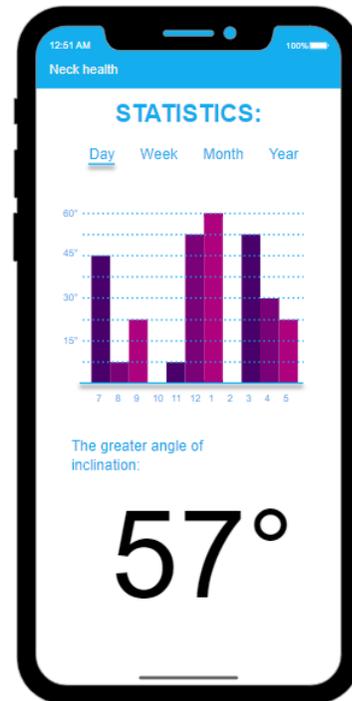


Fig. 8. Module for statistics of the device's tilt angles

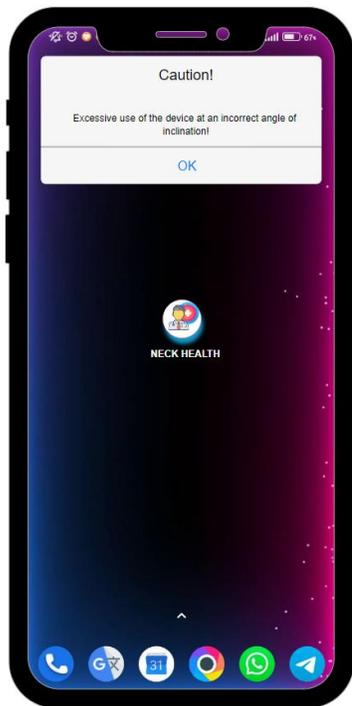


Fig. 9. Notification to the user about excessive use of the cell phone at an improper tilt angle

#### D. Notification of Tilt Angle

In this module, after the system detects an inadequate angle of inclination of the device and an excessive use of time, a notification is made. In Fig. 9 it can be seen that the application notifies the user of the excessive use of the device by means of a message: "Excessive use of the device with an incorrect angle of inclination", by means of this notification the user can become aware that he is taking a bad position of the mobile device allowing him to change it to safeguard his health.

#### IV. DISCUSSIONS

In this section, we show the differences of our research work in the various topics that coincide in those works that have been used as a basis for the development of the same.

In this section, we show the differences of our research work in the various topics that coincide in those works that have been used as a basis for the development of the same [1], [2], [4]–[8]. Indeed, several research studies have collected data on the use of various methodologies and recommendations to counteract the effects caused by this syndrome on users [1], [3], [9], [10]. However, the development of a system that allows the user to become aware of the moment in which the position adopted when using a mobile device is detrimental to their health is not performed. Therefore, this research project implements a system that allows the user to become aware of these bad postures when using mobile devices.

At present, mobile devices such as cell phones have become very popular, reaching 81% of acceptance by Americans [13]. They are also used to monitor the physical activities of their users through the use of various sensors [11], [12], [14], [19]. Among the sensors most commonly used in activity monitoring are the gyroscope and the accelerometer

because they provide accurate information [11], [14]–[16], [18], [20]. In this sense, the information provided by these sensors can be used to determine where the device is being used, in addition to knowing precisely how long the user is in that position.

#### V. CONCLUSIONS

In conclusion, it is possible to use the information of the tilt angle of the mobile devices through the use of the gyroscope and accelerometer sensors of the mobile devices. This information allowed the system to show the user the current tilt angle of the device. It is of utmost importance that the user is aware of the misuse of the device, since this allows the user to change position quickly in order to avoid falling into the text neck syndrome. Also, having a record of the activity that has taken place over time, greatly helps the user to keep track of their progress. Finally, the notifications provided by the system allow the user to have a timelier knowledge of their incorrect position during the use of the device, allowing it to make an immediate change in its position.

#### REFERENCES

- [1] M. T. Ansari and S. Ghosh, "CONCEPT OF TEXT NECK SYNDROME AND ITS MANAGEMENT THROUGH AYURVEDA," *International Ayurvedic Medical Journal*, vol. 9, no. 2, pp. 447–452, Feb. 2021, doi: 10.46607/iamj2009022021.
- [2] M. A. Kulkarni, N. Lakhwani, N. Prajapati, D. Prajapati, B. Modi, and P. Muni, "Effect of providing awareness regarding the text neck syndrome in young adults," *Int J Community Med Public Health*, vol. 9, no. 2, p. 831, Jan. 2022, doi: 10.18203/2394-6040.ijcmph20220247.
- [3] D. Jain, S. Jawade, and N. Chitale, "Effectiveness of Progressive Resisted Exercise along with Conventional Exercise and Conventional Exercise Program alone in Subjects with Text Neck Syndrome," *J Pharm Res Int*, pp. 536–542, Dec. 2021, doi: 10.9734/jpri/2021/v33i59b34412.
- [4] A. Kaur and S. Makker, "A Study to Assess the Prevalence of Text Neck Syndrome and Quality of Sleep among Smartphone Users in Selected Colleges of District Ludhiana, Punjab," *Int J Health Sci Res*, vol. 11, no. 9, pp. 49–54, Sep. 2021, doi: 10.52403/ijhsr.20210907.
- [5] S. Kumari, R. Kumar, and D. Sharma, "Text Neck Syndrome: The Pain of Modern Era," *Int J Health Sci Res*, vol. 11, no. 11, pp. 161–165, Nov. 2021, doi: 10.52403/ijhsr.20211121.
- [6] M. Kamalakannan, R. Rakshana, and R. Padma Priya, "Estimation and prevention of text neck syndrome among smart phone users," *Biomedicine (India)*, vol. 40, no. 3, 2020, doi: 10.51248/v40i3.30.
- [7] M. Gałczyk, W. Kułak, and A. Zalewska, "Is the use of mobile phones good for your neck? Text neck syndrome as an awareness of the existing threat – literature review," *Medycyna Ogólna i Nauki o Zdrowiu*, vol. 26, no. 3, pp. 240–243, Sep. 2020, doi: 10.26444/monz/126848.
- [8] P. P. Samani, N. A. Athavale, A. Shyam, and P. K. Sancheti, "Awareness of text neck syndrome in young-adult population," *Int J Community Med Public Health*, vol. 5, no. 8, p. 3335, Jul. 2018, doi: 10.18203/2394-6040.ijcmph20183057.
- [9] D. David, C. Giannini, F. Chiarelli, and A. Mohn, "Text neck syndrome in children and adolescents," *Int J Environ Res Public Health*, vol. 18, no. 4, 2021, doi: 10.3390/ijerph18041565.
- [10] O. Soyer and Z. Ü. Akarımak, "The Effect of Postural Correction and Exercise on Neck Pains in Cell Phone Users," *Turk Osteoporoz Dergisi*, vol. 26, no. 2, pp. 81–91, Aug. 2020, doi: 10.4274/tod.galenos.2019.76094.
- [11] A. Prasad, A. K. Tyagi, M. M. Althobaiti, A. Almulihi, R. F. Mansour, and A. M. Mahmoud, "Human activity recognition using cell phone-based accelerometer and convolutional neural network," *Applied Sciences (Switzerland)*, vol. 11, no. 24, Dec. 2021, doi: 10.3390/app112412099.

- [12] M. Y. Esas and F. Latifoğlu, "Motion Type Verification Studies Using Accelerometer Sensor Data With Local Mean Decomposition," *Academic Perspective Procedia*, vol. 2, no. 3, pp. 1051–1057, Nov. 2019, doi: 10.33793/acperpro.02.03.117.
- [13] M. A. Emberson, A. Lalonde, D. Wang, D. J. Mcdonough, W. Liu, and Z. Gao, "Effectiveness of Smartphone-Based Physical Activity Interventions on Individuals' Health Outcomes: A Systematic Review," *BioMed Research International*, vol. 2021. Hindawi Limited, 2021. doi: 10.1155/2021/6296896.
- [14] A. J. Casson, A. Vazquez Galvez, and D. Jarchi, "Gyroscope vs. accelerometer measurements of motion from wrist PPG during physical exercise," *ICT Express*, vol. 2, no. 4, pp. 175–179, Dec. 2016, doi: 10.1016/j.icte.2016.11.003.
- [15] A. Cieccko, G. Grunwald, K. Kowalski, D. Tanajewski, and M. Dobek, "Analysis of the Accuracy and Usefulness of MEMS Chipsets Embedded in Popular Mobile Phones in Inertial Navigation," in *IOP Conference Series: Earth and Environmental Science*, Mar. 2019, vol. 221, no. 1. doi: 10.1088/1755-1315/221/1/012070.
- [16] A. Pranav, "International Research Fibrillation Detection using Accelerometer and Gyroscope of a Smartphone," 2018. [Online]. Available: [www.ijtsrd.com](http://www.ijtsrd.com)
- [17] C. Fennell, E. L. Glickman, A. Lepp, J. D. Kingsley, and J. E. Barkley, "The Relationship between Cell Phone Use, Physical Activity, and Sedentary Behavior in United States Adults above College-age," *International Journal of Human Movement and Sports Sciences*, vol. 6, no. 4, pp. 63–70, Dec. 2018, doi: 10.13189/saj.2018.060401.
- [18] V. v. Soshenko et al., "Nuclear Spin Gyroscope based on the Nitrogen Vacancy Center in Diamond," *Phys Rev Lett*, vol. 126, no. 19, May 2021, doi: 10.1103/PhysRevLett.126.197702.
- [19] M. Cimera and A. Voloshin, "Validation of smartphone sway analysis for fall prevention," *Applied Sciences (Switzerland)*, vol. 11, no. 22, Nov. 2021, doi: 10.3390/app112210577.
- [20] C. Jia and B. L. Evans, "Online camera-gyroscope autocalibration for cell phones," *IEEE Transactions on Image Processing*, vol. 23, no. 12, pp. 5070–5081, Dec. 2014, doi: 10.1109/TIP.2014.2360120.
- [21] S. Sanchana, A. P. Manjari, and B. N. Jyothi, "Application Development for Real-Time Location Tracking for Underwater Vehicles Using Low-Cost GPS with GSM," in *2021 International Conference on Wireless Communications, Signal Processing and Networking, WiSPNET 2021*, Mar. 2021, pp. 108–112. doi: 10.1109/WiSPNET51692.2021.9419421.
- [22] U. Can and B. Alatas, "A new direction in social network analysis: Online social network analysis problems and applications," *Physica A: Statistical Mechanics and its Applications*, vol. 535. 2019. doi: 10.1016/j.physa.2019.122372.
- [23] S. Biroğul and H. B. Gültekin, "Importance of Business Intelligence Solution on Decision-Making Process of Companies," *International Journal of Applied Mathematics, Electronics and Computers*, pp. 86–86, 2016, doi: 10.18100/ijamec.266141.
- [24] A. E. Mogrovejo-lazo, Celio Froilán Andrade-Cordero, and Héctor Alejandro Espinoza-Pillaga, "Inteligencia de Negocios con enfoque estratégico en el sector comercial de la ciudad de Cuenca-Ecuador," *REVISTA CIENTIFICA DOMINIO DE LAS CIENCIAS*, vol. 7, no. 3, 2021.
- [25] A. Nasution, B. Efendi, and I. Kamil Siregar, "PELATIHAN MEMBUAT APLIKASI ANDROID DENGAN ANDROID STUDIO PADA SMP NEGERI 1 TINGGI RAJA," *Jurdimas (Jurnal Pengabdian Kepada Masyarakat)* Royal, vol. 2, no. 1, 2019, doi: 10.33330/jurdimas.v2i1.321.
- [26] C. Khawas and P. Shah, "Application of Firebase in Android App Development-A Study," *Int J Comput Appl*, vol. 179, no. 46, 2018, doi: 10.5120/ijca2018917200.
- [27] A. V. B. Dr.Amita Goel and Nidhi Sengar, "Android Based Instant Messaging Tool Using Firebase as Backend," *International Journal for Modern Trends in Science and Technology*, vol. 6, no. 12, pp. 198–201, Dec. 2020, doi: 10.46501/ijmst061238.
- [28] E. A. W. Sanad, "Pemanfaatan Realtime Database di Platform Firebase Pada Aplikasi E-Tourism Kabupaten Nabire," *Jurnal Penelitian Enjiniring*, vol. 22, no. 1, 2019, doi: 10.25042/jpe.052018.04.
- [29] R. Mallik, A. P. Hazarika, S. Ghosh Dastidar, D. Sing, and R. Bandyopadhyay, "Development of An Android Application for Viewing Covid-19 Containment Zones and Monitoring Violators Who are Trespassing into It Using Firebase and Geofencing," *Transactions of the Indian National Academy of Engineering*, vol. 5, no. 2, 2020, doi: 10.1007/s41403-020-00137-3.

# An Early Warning Model for Intelligent Operation of Power Engineering based on Kalman Filter Algorithm

Haopeng Shi, Xiang Li, Pei Sun, Najuan Jia, Qiyang Dou\*

State Grid Gansu Electric Power Company, Pingliang Power Supply Company, Pingliang Gansu 744000, China

**Abstract**—The accurate early warning of intelligent operation of power engineering can find the abnormal operation of substation equipment in time and ensure the safe operation of substation equipment. Thus, an early warning model for intelligent operation of power engineering based on Kalman filter algorithm is constructed. In this model, the noise elimination method of substation equipment inspection image based on particle resampling filter algorithm is introduced. After removing the noise information of operation situation inspection image of substation equipment, the gradient direction histogram feature, lab color space feature and edge contour feature in the image are extracted by the multi-feature extraction method for intelligent operation of power engineering based on multi-feature fusion. These features are combined to form the feature description set of equipment operation situation. The feature description set is used as the identification attribute set of the anomaly identification and early warning model for intelligent operation of electric power engineering based on Kalman filter algorithm to complete the anomaly identification and early warning of equipment operation situation. The test shows that when the model is used to observe the temperature change trend of the top layer of the transformer, the temperature error is very small, and the early warning accuracy for the abnormal temperature of the top layer of the transformer is very high, so the abnormal operation of the substation equipment can be found in time.

**Keyword**—Kalman filter; power engineering; intelligent operation; early warning model; image denoising; feature extraction

## I. INTRODUCTION

With the development of industrial revolution, electric power plays an important role in the development of human society. Problems in urban power supply system will cause serious consequences for people's daily activities and even the whole society. Therefore, ensuring the safe and stable operation of power system has become an extremely important part of the national strategic energy security system. To realize the safe and reliable operation of the power grid, it is necessary to continuously improve the automation level of the substation, so as to achieve the purpose of reasonable allocation of power supply equipment and effective supervision and management of the substation equipment [1].

The development of substation automation is mainly affected by the following aspects. First, the construction of substation automation system is inseparable from information technology. The development of information technology makes the information source and amount faced by power enterprises continue to grow. Therefore, enterprises also put

forward relatively high requirements for the work efficiency and response speed of the processing system [2]. Secondly, the automation level of the substation mainly involves fault detection and fault isolation. On the one hand, the wide distribution of low-voltage substation has resulted in the large number and scattered nodes of low-voltage substation. Therefore, the staff must upload patrol inspection data in real time, accurately and efficiently. On the other hand, due to the long-term exposure of power equipment to the field, it often bears the effects of normal mechanical load and power load, as well as external forces such as lightning strike, pollution, strong wind, earthquake, flood, landslide and bird damage, and may even be endangered by the theft of power equipment by some criminals. These factors will cause aging, oxidation and corrosion of various components in the substation. If they are not found and eliminated in time, the existing hidden dangers will develop into faults, thus posing a threat to the security and stability of the power system [3]. Therefore, in order to fundamentally ensure the safe and stable operation of the power grid and the safe supply of power, it is necessary to carry out regular and irregular inspection of each substation, timely find hidden dangers, prevent them before they happen, and reduce the failure rate of power equipment to the lowest.

At present, the traditional manual inspection, video monitoring, comprehensive maintenance vehicle and other methods are mainly used for substation inspection in China's power system. For some enterprises, the cost is too high, the operation is difficult to implement, and it is not easy to promote. At the same time, it is difficult to eliminate the impact of human factors in the inspection work. Therefore, how to absorb the modern management experience of foreign advanced power enterprises and make full use of advanced mobile technology in the construction of intelligent inspection system for substation has become very important.

## II. JOURNALS' REVIEWED

Xie, S studied the intelligent inspection technology of substation electrical equipment based on 5G. Based on 5G communication technology, intelligent patrol robots, video surveillance, AR glasses, mobile patrol APP and other terminals are integrated to realize all-round intelligent patrol of the substation, with poor early warning effect [4]. Yang Qiong designed an intelligent patrol inspection system for substation, which introduced GPS and PDA technology, and was characterized by the storage of equipment information in the upper management system and all management of equipment, defect information, historical data and patrol inspection. Patrol inspectors could complete the information collection of equipment only by arriving near the corresponding patrol

inspection equipment. After the patrol inspection, the corresponding equipment information should be transmitted to the background database for storage through a certain communication mode. The comprehensive cost of this system was relatively high [5]. Zhang F. researched the inspection method combining PDA and RFID. This inspection method was a non-contact automatic identification technology, which could read and identify the electronic data stored in the card without contact. The reader / writer emitted energy in an area to form an electromagnetic field. When the RF tag passed through the area, it detected the signal of the reader / writer and sends the stored data. The reader / writer received the signal sent by the RF tag, decoded and checked the accuracy of the data to complete the identification, so as to achieve the purpose of patrol inspection [6].

Compared with other methods, RFID technology has the characteristics of non-contact identification, high-speed identification, multi-target simultaneous identification, and strong confidentiality. It is widely used in vehicle identification and production process control, but the immunity of radio frequency technology needs to be optimized. In order to discover the abnormal operation of substation equipment in time and realize the efficient and intelligent operation early warning of power engineering. On the basis of previous studies, this paper establishes an intelligent operation early warning model of electric power engineering based on Kalman filter algorithm, which is mainly used for intelligent patrol inspection of electric power engineering, in order to provide some help for timely early warning of abnormal conditions found in the patrol inspection process.

### III. EARLY WARNING MODEL FOR INTELLIGENT OPERATION OF POWER ENGINEERING

The early warning of intelligent operation of power engineering needs to be completed by using intelligent technology. At present, the application of intelligent inspection robot is no longer strange in the field of power engineering. The intelligent inspection robot has replaced the traditional manual inspection mode. The early warning model for intelligent operation of power engineering based on Kalman filter algorithm constructed in this paper belongs to one of the core technologies applied to the equipment of intelligent inspection robot. Before introducing the specific application technology of the model, the structure of the intelligent inspection robot is analyzed. The structure diagram of its operation mode is shown in Fig. 1.

The intelligent inspection robot system consists of a base station layer and a robot mobile station. The base station layer receives and processes patrol inspection data through its database, data processing and video monitoring modules, which is equivalent to a monitoring background. It also has the functions of image processing and pattern recognition, which can automatically identify equipment defects and automatically warn. The communication layer is divided into two modules: the wireless bridge base station and the wireless bridge mobile station. It provides data transmission channels for the base station and the robot mobile station through the wireless network transmission protocol. Wireless communication is used between the mobile robot and the monitoring background

[7]. In addition, when the robot loses power, it can automatically return to the charging room for self-charging. The model in this paper is mainly installed on the intelligent inspection robot equipment to help the robot find the abnormal situation of intelligent operation of power engineering in real time and give real-time warning. The intelligent operation warning problem of power engineering studied in this paper is mainly to identify and warn the operation situation of substation equipment.

#### A. Noise Elimination Method for Inspection Image of Substation Equipment based on Particle Resampling Filter Algorithm

When the intelligent inspection robot performs the inspection task of operation situation of substation equipment, the collected infrared image is affected by external factors, so there is inevitably noise information, which directly affects the image quality [8]. For this reason, this paper uses the noise elimination method of inspection image of substation equipment based on particle resampling filter algorithm to remove the noise information of inspection image and optimize the image quality [9].

The particle filter algorithm uses the large number theorem to solve the nonlinear non Gaussian estimation problem in Bayesian estimation through Nonparametric Monte Carlo, which is applicable to any nonlinear non Gaussian random problem that can be expressed in state space [10]. The posterior probability density of noise particles is estimated through a group of observed random noise samples in the state space of noise particles in the inspection image, and the mean value of noise samples is used to replace the integral operation to obtain the minimum variance of noise filtering effect.

The infrared image state equation and noise observation equation of operation situation inspection of substation equipment are modeled as follows:

$$\begin{cases} y_h = g(y_h, u_h) \\ l_h = k(y_h, m_h) \end{cases} \quad (1)$$

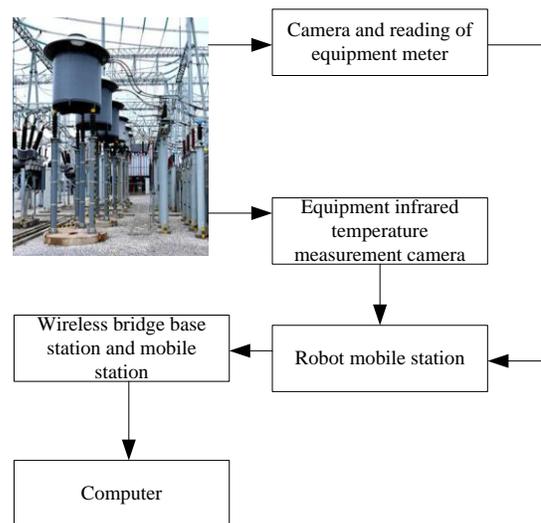


Fig. 1. Structure diagram of operation mode of intelligent inspection robot.

Where,  $u_h$  is the actual noise level during patrol inspection image acquisition;  $m_h$  is the measurement noise level of patrol inspection image;  $g(\cdot)$  is the transfer function of patrol inspection image quality;  $k(\cdot)$  is the measurement function of patrol inspection image quality.  $y_h$  is a group of filter state values of weighted particle with conditional distribution obtained by Monte Carlo simulation sampling at time  $h$ . Each noise particle uses its gray value at the patrol inspection image position as the characteristic value, and  $l_h$  is the observation value of noise particle filter at time  $h$ .

Through the probability density function  $q(y_h | l_{1:h-1})$  of the spatial state of the noise particle swarm optimization system at time  $h-1$  based on the Chapman-Kolmogorov equation, the state of noise information in the patrol inspection image of substation equipment's operation situation at time  $h$  is observed:

$$q(y_h | l_{1:h-1}) = \int q(y_h | y_{h-1}) q(y_{h-1} | l_{1:h-1}) dy_{h-1} \quad (2)$$

Where,  $l_{1:h-1}$  is the noise observation value from 1 to  $h-1$ .

The Bayesian formula is used to derive the prior probability  $q(y_h | l_{1:h-1})$  and the posterior probability  $q(y_h | l_{1:h})$  from the noise observation value at time  $h$ . According to the law of large numbers, when the number of noise particles is very large, the particle filter is approximate to the posterior probability of the patrol inspection image quality state [11]. Namely:

$$q(y_h | l_{1:h}) \approx \frac{1}{M} \sum_{j=1}^M \varpi_h^j \beta(y_h - y_{1:h}^{j(n,f)}) \quad (3)$$

Where  $M$  is the number of noise particles;  $\varpi_h^j$  is the weight of noise particle  $j$  at time  $h$ ;  $\beta$  is a Dirac function;  $y_{1:h}^{j(n,f)}$  is the  $j$ -th noise particle gray value located at the patrol inspection image  $(n, f)$  of operation situation of substation equipment from 1 to  $h$ . Generally, the noise particle set cannot be directly sampled from the posterior probability. The prior density that is easy to realize is selected as the importance density function, and the particles with uniform distribution are optimized through maximum likelihood estimation, so that the optimized particle distribution is closer to the posterior probability density. Only a small number of noise particles can achieve high estimation accuracy, thus reducing the amount of calculation [12]. Then the weight is updated to:

$$\varpi_h = \varpi_h^j q(y_h | l_{1:h}) \quad (4)$$

Weight normalization:

$$\hat{\varpi}_h = \frac{\varpi_h^j}{\sum_{j=1}^M \varpi_h^j} \quad (5)$$

The optimal image quality of patrol inspection of substation equipment's operation situation is output:

$$\hat{y}_h = \sum_{j=1}^M \hat{\varpi}_h^j y_h \quad (6)$$

After several iterations of particle algorithm, only a small number of effective noise particles have non-zero important weights, and most of the important weights of noise particles tend to zero. Therefore, the noise cannot be effectively eliminated in the process of infrared image denoising. In order to prevent the weight degradation of noise particles, resampling method is adopted to reduce the impact of noise on the image quality of patrol inspection of substation equipment's operation situation to a certain extent [13].

The main idea of resampling is to remove the noise particles with small weight, retain and copy the noise particles with large weight, and sample the particles with large weight for many times, so as to increase the chance of noise elimination for the particles, and sample less for the particles with small weight [14]. Firstly,  $m$  uniformly distributed random numbers  $\{\varepsilon_i; i=1, 2, \dots, m\}$  are generated in the interval  $[0, 1]$ , and then the weights of noise particles are accumulated:

$$d_h = \sum_{j=1}^h \hat{\varpi}_h \quad (7)$$

Where,  $h=1, 2, \dots, m$ ;  $d_h$  is the cumulative value of noise particle weight.

The resampled noise particles only account for a small part of the whole particle swarm. Such particles can no longer effectively describe the posterior probability distribution of patrol inspection image quality status. As each particle is sampled independently, the process includes cycle and comparison operations, which increases the computational complexity of particle filter for patrol inspection image of substation equipment's operation situation [15].

Resampling makes the noise particles with higher weight be sampled for many times, and the sampling results contain many repeated noise particles, thus losing the diversity of particles and reducing the filtering performance. In order to solve the problem of resampling dilution, the effective particle number threshold is used to optimize the filtering effect of patrol inspection image of substation equipment's operation situation.

Suppose  $\{\hat{\omega}_h^1, \hat{\omega}_h^2, \dots, \hat{\omega}_h^M\}$  is the set of normalized weights of noise particles, and the sample variance of particle weights is:

$$U(\hat{\omega}_h) = \frac{1}{M} \sum_{j=1}^M [\hat{\omega}_h - \text{mean}(\hat{\omega}_h)]^2 \quad (8)$$

Where  $U(\cdot)$  is the variance function;  $\text{mean}(\cdot)$  is the mean function.

Effective particle number threshold  $M_{\text{eff}}(\hat{\omega}_h)$  of the degradation degree of noise particles is measured:

$$M_{\text{eff}}(\hat{\omega}_h) = \frac{1}{\sum_{j=1}^N (\hat{\omega}_h)^2} \quad (9)$$

In this way, the effective noise particles are classified according to the weight value, which effectively reduces the complexity of the algorithm, and the random method increases the diversity of noise particles. In order to avoid too small noise particles in the patrol inspection image of substation equipment's operation situation, the lower limit of noise variance is set as  $\alpha_{\min}$ , and the maximum noise variance is set as  $\alpha_{\max}$ . The attenuation factor  $\eta$  is used to estimate the noise reduction rate. When the noise in the patrol inspection image of substation equipment's operation situation is small, a small number of particles can be used to describe the noise distribution. When the noise is large, the sampling range of particles is expanded and the number of particles is increased. Sigmoid function is used to express the relationship between the number of noise particles  $M_h$  and uncertainty measure  $o_h$  at time  $h$ .

$$M_h = \frac{2M_{\text{eff}}(\hat{\omega}_h)}{1 + \exp(-\eta o_t + \eta \cdot \eta_{\min})} \quad (10)$$

Among them, the uncertainty measures  $o_h = \theta_t^x \theta_t^z$ ,  $\theta_t^x$  and  $\theta_t^z$  are the actual noise and observation noise in the patrol inspection image of substation equipment's operation situation at time  $h$ . In this way, the noise particles with smaller weight are discarded and replaced by the noise particles with larger weight for multiple noise elimination. The noise particles with larger weight are erased after the patrol inspection image is resampled to complete the noise elimination of the patrol inspection image of substation equipment's operation situation [16].

### B. Multi-Feature Extraction Method for Intelligent Operation of Power Engineering based on Multi-Feature Fusion

In order to identify the abnormal situation in the inspection process of substation equipment's operation situation, it is

necessary to extract the image features of substation equipment's operation situation after de-noising in Section IIIA. The image features of inspection are very important in the process of substation equipment's operation situation awareness. Therefore, this paper will comprehensively consider the edge, gradient and color features in the infrared image during intelligent inspection of substation equipment. A multi-feature extraction method for intelligent operation of power engineering based on multi-feature fusion is proposed. The feature description set of equipment operation situation is composed of gradient direction histogram (HOG) feature, lab color space and edge contour, which is used as the identification attribute set of the anomaly identification and early warning model of intelligent operation of power engineering based on Kalman filter algorithm.

1) *Gradient feature extraction*: The gradient of patrol inspection image of substation equipment's operation situation includes image edge contour and texture information, which can be used for image analysis and recognition. In the process of extracting the image features of substation equipment's operation situation inspection, this paper simplifies the calculation process of HOG features, improves the calculation efficiency, and can better describe the gradient features of substation equipment's operation situation. In the 5\*5 cell, 9 bin histograms are used to calculate the gradient information of these 25 pixels. That is, the 360 ° gradient direction of the cell is divided into nine direction intervals, and each pixel in the cell is weighted projected in the histogram with the gradient direction. The weight is the gradient amplitude, and the amplitude of the histogram in each direction forms an eigenvector [17].

The gradient of all pixels in the patrol inspection image of substation equipment's operation situation after noise removal is calculated, and the gradient amplitude is:

$$F(x, y) = F_x(x, y) + F_y(x, y) \quad (11)$$

Where,  $F_x(x, y)$  and  $F_y(x, y)$  are the gradient amplitudes in direction  $x$  and  $y$  of each pixel in the inspection image of the substation equipment's operation situation after noise removal.

The gradient direction is:

$$\mu(x, y) = \tan^{-1} \frac{F_x(x, y)}{F_y(x, y)} \quad (12)$$

The weighted projection of each pixel in the cell within the uniform interval in the gradient direction  $\left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$  is calculated as:

$$F_k(x, y) = \sum_{\mu(x, y) \in c_k} F(x, y) \quad (13)$$

Where,  $F_k(x, y)$  is the cumulative value of gradient amplitude in different gradient directions in the cell;  $c_k$  represents the range of different gradient directions.

2) *Color feature extraction:* After de-noising, the component L in the lab color space of the inspection image of substation equipment's operation situation expresses the human eye's perception of brightness. The output color scales of components a and b are more uniform and balanced. Compared with RGB and CMYK color models, the lab space has a broader color gamut and is independent of physical equipment.

Therefore, in order to preserve the wide color gamut and rich colors as much as possible, and better quantify the colors, this paper uses the lab color model as the color feature, that is, the patrol inspection image of substation equipment's operation situation after noise removal is transformed into the lab color space model, and the lab color space is divided into three feature vector sets.

3) *Edge profile features:* The edge of the inspection image of substation equipment's operation situation after noise removal refers to the area where the local gray level of the image changes significantly. It is the most basic feature of the image and contains useful information for identification. Therefore, this paper extracts the direction, first-order and second-order differentiation of the image as the edge contour feature vector set of the de-noising patrol inspection image of substation equipment's operation situation [18]. Among them, the first-order and second-order differential are realized by Sobel differential operator.

To sum up, the extracted feature vectors of inspection image of substation equipment's operation situation after noise removal are used to form a feature descriptor. Each feature vector is a feature channel, and each feature channel is a matrix block with the same size as the image.

### C. Anomaly Identification and Early Warning Model for Intelligent Operation of Power Engineering based on Kalman Filter Algorithm

The characteristic information of the patrol inspection image of substation equipment's operation situation extracted in Section IIIB is used as the identification attribute set of the intelligent operation anomaly identification and early warning model of power engineering based on the Kalman filter algorithm. The Kalman filter algorithm mainly includes two processes: prediction and correction, that is, observation and update. The observation process mainly uses the time updating equation to establish a prior estimate of the current substation equipment's operation situation, so as to calculate the current state variables and error covariance estimates in time, and construct a prior estimate for the next time state; In the correction process, a posteriori estimate of the current state of substation equipment's operation situation is established based on the prior estimate of the prediction process and the current measurement variables by using the measurement update

equation through feedback. This process is called the prediction correction process.

In order to apply the Kalman filter algorithm to the intelligent operation early warning of power engineering, it is necessary to construct the description equation and measurement equation of substation equipment's operation situation based on the Kalman filter algorithm, and then establish the real-time optimal estimation model of substation equipment's operation situation [19]. Then the equation describing the operation situation of substation equipment is:

$$\alpha_{oil,h} = \Omega \alpha_{oil,h-1} + B_2 \begin{bmatrix} \alpha_{oil,h} \\ N \end{bmatrix} + V_{h-1} \quad (14)$$

Where,  $\alpha_{oil,h}$  is the operation situation of substation equipment at time  $h$ ;  $\Omega$  is the characteristic information of patrol inspection image of substation equipment's operation situation extracted in subsection 2.2;  $B_2$  is the gain of control input of substation equipment;  $V_{h-1}$  is the process excitation noise, which is generally considered to obey the normal white noise and does not change with time. It represents the observation error of substation equipment's operation situation between  $h-1$  and  $h$ ;  $N$  is the observation times of operation situation of substation equipment;  $\alpha_{oil,h-1}$  is the operation situation of substation equipment at time  $h-1$ .

The observation equation of the operation situation of substation equipment at time  $h$  can be expressed as:

$$L_K = H_k \alpha_{oil,h} + W_K \Omega \quad (15)$$

Where,  $H_k$  is the gain of actual variable  $\alpha_{oil,h}$  of substation equipment's operation situation to the observation variable  $L_K$  of substation equipment's operation situation;  $W_K$  is the change range of operation situation of substation equipment.

After determining the state equation and observation equation of the operation situation of the substation equipment, the Kalman filter algorithm estimates the operation situation of the substation equipment, which can carry out the two main processes of the Kalman filter: "time update (observation)" and "state update (correction)". Through repeated update and correction, the most accurate results can be obtained to realize the observation of the operation situation of the substation equipment. Firstly, it should establish the time update equation for the operation of substation equipment:

$$\hat{\alpha}_{oil,h}^- = \Omega \hat{\alpha}_{oil,h-1}^+ + B_2 \begin{bmatrix} \alpha_{amb,h} \\ N \end{bmatrix} \quad (16)$$

$$Q_K^- = (1 - B_1)^2 \cdot Q_K^+ + P \quad (17)$$

Where, "-" stands for a priori and "+" stands for a posteriori.  $\hat{\alpha}_{oil,h}^-$  is the prior state estimation of step  $h$  when the operation situation of the substation equipment before step  $h$  is known, that is, the prior state estimation of time  $h$  using time  $h-1$ .  $\hat{\alpha}_{oil,h-1}^+$  is a posteriori state estimation when the measurement variable  $L_K$  is known in the operation situation of substation equipment, and it is also the optimal estimation result of the state at  $h-1$ .  $Q_K^-$  is the covariance of the error of prior estimation, and  $P$  is the process error of substation equipment's operation situation estimation. In the updating process, the state estimation is optimized by using the prior estimates and observations of the current state, which is called a posteriori state estimation.  $Q_K^+$  is the covariance of the error of the calculated posterior estimate.

The verification equation for the estimation results of substation equipment's operation situation is:

$$\hat{\alpha}_{oil,h}^+ = \Omega \hat{\alpha}_{oil,h}^- + F_h (L_K - H_k \hat{\alpha}_{oil,h}^-) \quad (18)$$

$$F_h = \frac{Q_K^-}{Q_K^- + S} \quad (19)$$

$$Q_K^+ = Q_K^- (1 - F_h) \quad (20)$$

In Eq. (18) to (20),  $S$  is the variance of the operation situation estimation error of the measured substation equipment. A posteriori estimate  $\hat{\alpha}_{oil,h}^+$  is composed of a linear combination between a priori estimate  $\hat{\alpha}_{oil,h}^-$  and the observation variable  $L_K$  of substation equipment's operation situation;  $F_h$  is the Kalman gain, whose function is to minimize the posterior estimation error covariance  $Q_K^+$  to ensure that the recursion can be carried out continuously. The magnitude of the residual value reflects the inconsistency between the observed value and the actual value. The greater the residual value is, the greater the inconsistency is. Otherwise, it is true. Eq. (17) to (20) constitute the five processes required for Kalman filter iterative observation. In the iterative process, equations (18) and (20) need to feed back the obtained posterior results to equations (17) and (18) in order to update the information of each step. In this way, the model can realize the real-time estimation of the operation situation of substation equipment and the real-time early warning of intelligent operation of power engineering.

The early warning program is mainly completed by the alarm device. The circuit of the alarm device is composed of high decibel active alarm. Because the driving ability of the single chip microcomputer is not enough, this circuit uses NPN triode to drive the alarm. The LED display circuit consists of two decoders 741138, eight row drivers 4953 (each chip

controls two rows), eight column drivers (each chip controls eight columns), and sixteen 1588 common anode diode lattice modules. Therefore, the LED display is a 1664 dot matrix. Four Chinese characters can be displayed at the same time. In normal state, the screen displays yellow "normal operation" and red "fault operation" in case of fault, accompanied by alarm sound [20]. The alarm flow of this device is shown in Fig. 2.

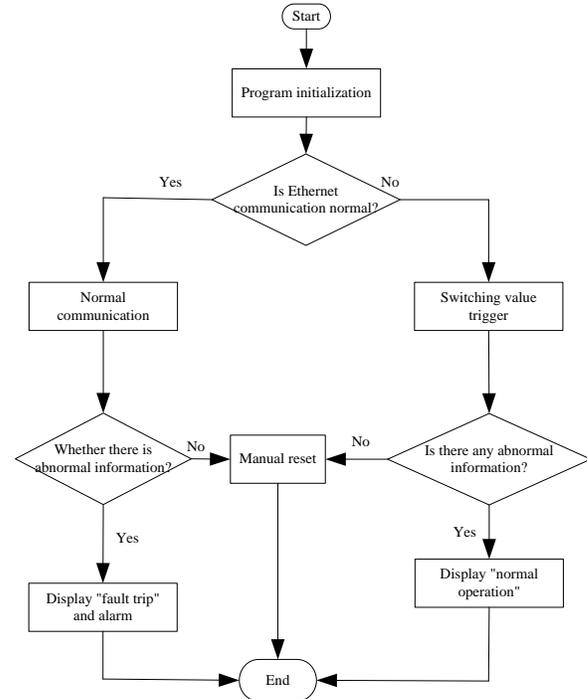


Fig. 2. Alarm process.

As shown in Fig. 2, the alarm device will first judge whether its Ethernet communication is normal. If the normal communication represents that the device can operate normally, it will judge whether there is any abnormal information touch device; if yes, it will display "fault operation" and sound light alarm. If there is no abnormality, it will display "normal operation" and reset manually.

#### IV. RESULTS

The model is installed on an intelligent substation inspection robot to test the application effect of the model. The test content is mainly divided into the processing effect, situation recognition and early warning effect of intelligent inspection image of substation equipment's operation situation,.

##### A. Analysis on Processing Effect of Intelligent Patrol Inspection Image of Substation Equipment's Operation Situation

As shown in Fig. 3(a) and Fig. 4(a), when the model in this paper inspects the operation situation of transformers and insulators in substations, there are different degrees of noise information in the captured intelligent inspection image of operation situation. The effect pictures of this model after noise removal are shown in Fig. 3(b) and Fig. 4(b).

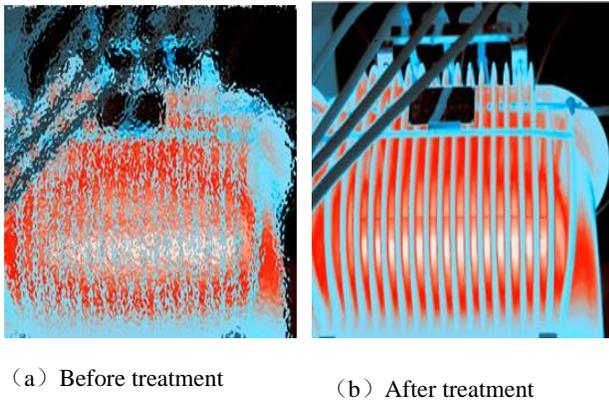


Fig. 3. Intelligent inspection image of transformer operation situation.

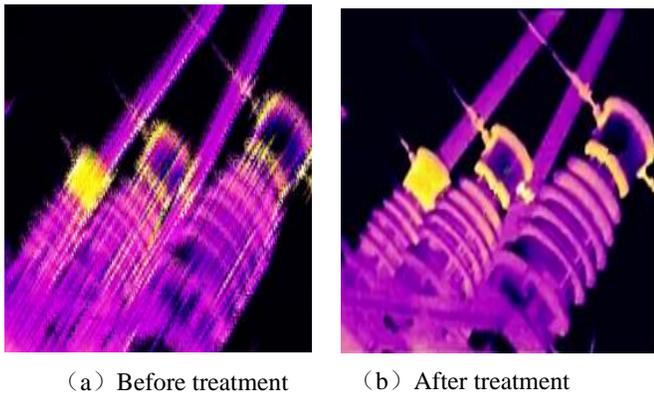


Fig. 4. Intelligent inspection image of insulator operation situation.

From the analysis of Fig. 3 and Fig. 4, it can be seen that the model in this paper has a good denoising effect on the transformer and insulator images of the patrol substation when inspecting the operation situation of the transformer and insulator in the substation. From the visual point of view, the image definition after denoising is improved and the image details are more significant.

### B. Abnormal Situation Identification and Early Warning Effect

The model in this paper is used to monitor the top temperature of the transformer shown in Fig. 3 in real time. The actual value of the top temperature of the transformer is shown in Fig. 5, and the ambient temperature is shown in Fig. 6.

The Kalman filter algorithm in the model of this paper can use the new patrol information to continuously observe and modify the new state estimates, so it can observe the top temperature in real time. The initial noise state of the state equation can be obtained from the statistical value of the variance function, and the variance of the observation error can be obtained from the statistical value of the temperature sensor error. The system state equation is used to optimally estimate the state variable, that is, the top temperature. The recognition results are shown in Fig. 7.

As shown in Fig. 7, the difference between the identification result of the transformer's top temperature and

the actual value of the model in this paper is very small, which can well reflect the dynamic change of temperature. As long as the given system's initial value does not deviate too far from the real initial value, the Kalman filter algorithm can converge to the final value.

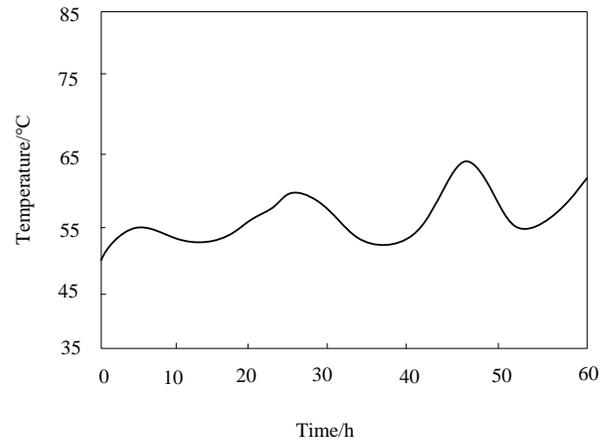


Fig. 5. Actual value of transformer top temperature.

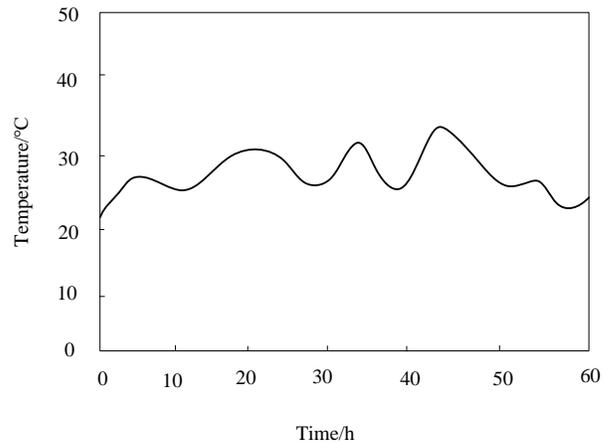


Fig. 6. Ambient temperature for intelligent operation of electric power engineering.

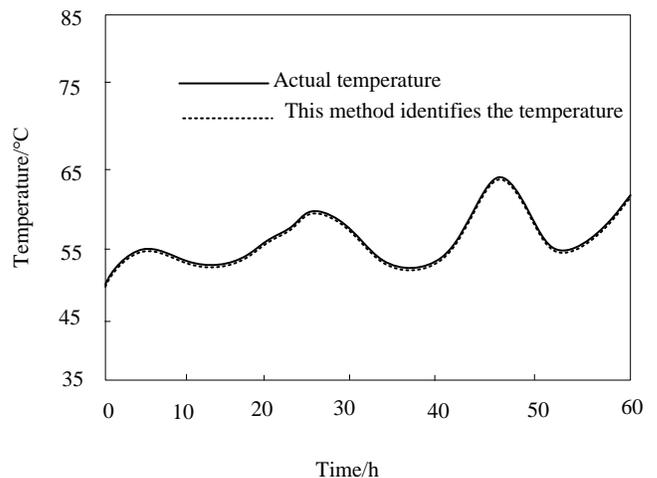


Fig. 7. The identification results of transformer top temperature by this model.

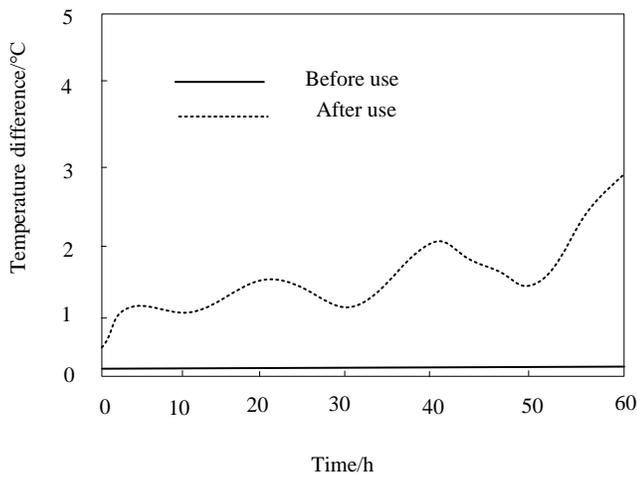


Fig. 8. The error variation of transformer top temperature identification results before and after the model is used in this paper.

Before and after the intelligent inspection robot uses the model in this paper, the error of the top temperature identification result of the transformer is shown in Fig. 8.

As shown in Fig. 8, when the model in this paper uses Kalman filter algorithm to observe the top temperature of

transformer, the temperature error is less than 0.5°. When the model in this paper is not used, the observation error of intelligent inspection robot on the top temperature of transformer is more than 2°. This shows that the model in this paper has high observation accuracy on the top temperature of transformer. The reason is that the Kalman filter algorithm can continuously predict and correct the top temperature of transformer, and the observation error is minimized. Therefore, this model can be well applied to the abnormal identification of transformer top temperature.

The abnormal early warning effect of the model in this paper after identifying the top temperature of transformer is test, and the results are shown in Table I.

By analyzing the data in Table I, it can be seen that after the model identifies the top-level temperature of the transformer, there is only a 1-minute delay in the early warning time for the abnormal state of the top-level temperature of the transformer on February 2, 2022, but the number of early warnings is consistent with the number of abnormal occurrences. The model in this paper has verified the early warning effect of intelligent operation of power engineering and can accurately give early warning.

TABLE I ABNORMAL EARLY WARNING EFFECT OF TOP LAYER TEMPERATURE OF MODEL TRANSFORMER

Abnormal occurrence date	Abnormal occurrence time	Warning time	Number of exceptions	Warning times
2022/1/23	10:58	10:58	5	5
2022/2/2	9:20	9:21	1	1
2022/3/25	8:01	8:01	2	2
2022/4/14	12:36	12:36	3	3
2022/5/8	20:25	20:25	5	5
2022/5/13	2:36	2:36	4	4
2022/6/1	17:25	17:25	1	1
2022/6/3	13:33	13:33	1	1
2022/6/4	12:45	12:45	1	1

## V. DISCUSSION

Based on the research content of this paper, at present, the intelligent inspection robot is mainly used in the intelligent operation inspection of power engineering. However, the intelligent inspection robot is not perfect, and it also has shortcomings in practical application:

1) *Meter data reading*: Patrol inspection robots are equipped with high-definition cameras that can be zoomed. According to the design idea, their advantages in reading meter data are far greater than the visual inspection of operation and maintenance personnel during manual patrol inspection. In the station, the installation position of some meters and meters is too high for manual vision to see the pointer, number and other contents in the meters and meters. In this case, the traditional inspection is conducted with the help of a telescope. The zoom camera of the intelligent inspection robot can not only shorten the meter interface several times the distance, save time and

effort, but also save the meter interface as a picture for later analysis. However, in practical application, it is found that this advantage cannot be fully exerted, which is mainly manifested in that when the surface of the high-voltage meter in the station or the camera of the inspection robot becomes dirty due to the accumulation of pollutants and impurities in the air, the camera simply cannot obtain a clear meter image. In addition, the camera also fails to focus. For the problem that the camera surface of the inspection robot is polluted, a self-cleaning device similar to the automobile wiper can be added in the subsequent improvement to properly solve it. As for the phenomenon of surface pollution, the problem of non-inspection robot itself can only be solved by manual cleaning during each power outage and maintenance of power grid equipment. Focusing failure is caused by the camera algorithm or the auto focusing technology adopted. It is recommended to configure the camera with active auto focusing mode, which

can greatly reduce the occurrence of focusing failure by combining the advantages of infrared ranging and ultrasonic ranging focusing methods and focusing mode based on image processing. At present, there are many researches and explorations on the research, development and application of inspection robot in the intelligent inspection system and the diversification of inspection functions, but the more advanced cutting-edge technologies are neglected in the hardware. High performance hardware equipment is more conducive to the advanced functions of the inspection robot.

2) *Infrared thermometry*: In the process of infrared temperature measurement, the layout of equipment in the station is complex and staggered, and the inspection robot is restricted by the fixed inspection path, positioning point and the traveling channel in the actual site designed in the system. It is unable to compare the maximum temperature and hot spot of the equipment from 360 ° directions flexibly like manual inspection, and there is inaccurate alignment. Therefore, the temperature information obtained is too large deviation from the actual situation, or even wrong. The problem of alignment and misalignment can be solved by optimizing the system and adding a distance monitoring unit. The function of the distance monitoring unit is to determine the reasonable distance between the tested equipment and the inspection robot. Only the equipment within a reasonable distance set in the system can be selected by the infrared camera, so as to effectively avoid the wrong selection of objects within the abnormal range such as the sum for temperature measurement in similar cases. However, in order to realize the full angle comparative temperature measurement like manual inspection, it is obviously impossible to realize it at the software level of the robot inspection monitoring system due to the factors such as the distribution of roads in the substation, the height of the inspection robot, the battery life and so on. Therefore, the infrared temperature measurement function cannot replace the accurate temperature measurement in the existing technical stage, and can only be used as a way of universal temperature measurement in a large area.

Therefore, the above two problems should be paid more attention in practical application.

## VI. CONCLUSION

With the rapid development of China's economy, higher requirements are put forward for the safe operation of power transmission network, power plant facilities and other infrastructure. Major power companies have invested a lot of manpower and material resources in the inspection and maintenance of lines and facilities. However, due to the limitations of technical conditions, there are many deficiencies in the power inspection link, such as low inspection efficiency and difficult management. At the same time, the original manual records and reports also greatly limit the modernization of power operation management. Intelligent inspection of substation equipment is applicable to the power supply facilities management department of the power supply company. It helps to reduce the workload of inspection

personnel and facilities management personnel and improve work efficiency. The research content of this paper is the early warning of intelligent operation of power engineering, which is simply the abnormal identification and early warning of substation equipment's operation situation. In order to accurately identify and early warning the abnormal state of intelligent operation of power engineering, this paper constructs an early warning model of intelligent operation of power engineering based on Kalman filter algorithm, and verifies its application value in experiments. However, in the experiment, the model in this paper has a one minute delay in identifying the abnormal operation situation of substation equipment. In the future research work, the application effect of the model will be gradually optimized to provide effective assistance for power equipment monitoring.

## ACKNOWLEDGMENTS

The study was supported by "State Grid Gansu Electric Power Company Management science and technology project support (522709220008)."

## REFERENCE

- [1] Liu, T. , Kuang, J. & Ge, W.(2021). A simple positioning system for large-scale indoor patrol inspection using foot-mounted INS, QR code control points, and smartphone. *IEEE Sensors Journal*, 21(4):4938-4948.
- [2] Zhou, Z., Du, Z. & Wang, W. (2019). Intelligent technology and application of live detection for substation operation and maintenance based on hierarchical distribution. *Dianli Xitong Baohu yu Kongzhi/Power System Protection and Control*, 47(1):150-157.
- [3] Ala, G. , Favuzza, S. & Mitolo, M. (2020). Forensic Analysis of Fire in a Substation of a Commercial Center *IEEE Transactions on Industry Applications*, 56(3):3218-3223.
- [4] Xie, S. (2021). Research and application of intelligent inspection technology for substation electrical equipment based on 5g Power Information and Communication Technology, 19 (12), 109-116.
- [5] Yang, Q.(2009). Design and implementation of substation intelligent patrol system [J] *Central China Power*, 2009, 22 (4): 35-37.
- [6] Zhang, F. (2013). PDA patrol management system based on RFID technology and its practical application *China Equipment Engineering* (2), 14-16.
- [7] Liu, H. , Wang, Y. & Chen, W. G.(2020). Anomaly detection for condition monitoring data using auxiliary feature vector and density-based clustering. *IET Generation, Transmission & Distribution*, 14(1):108-118.
- [8] Shi, Y., Ji, S.& Zhang, F. (2019). Multi-Frequency Acoustic Signal Under Short-Circuit Transient and Its Application on the Condition Monitoring of Transformer Winding. *IEEE Transactions on Power Delivery*, 34(4):1666-1673.
- [9] Zhang, J., Du, X. & Xiao, W. (2019).Condition Monitoring the Health Status of Forced Air Cooling System Using the Natural Frequency of Thermal Network. *IEEE Transactions on Power Electronics*, 34(11):10408-10413.
- [10] Liu, J. , Zhang, G. & Chen, Q. (2019). In situCondition Monitoring of IGBTs Based on the Miller Plateau Duration. *IEEE Transactions on Power Electronics*, 34(1):769-782.
- [11] Peng, L., Wang, Z. (2020). Grouping Capacitor Voltage Estimation and Fault Diagnosis With Capacitance Self-Updating in Modular Multilevel Converters. *IEEE Transactions on Power Electronics*, 36(2):1-1.
- [12] Zhuang, T., Ren, M.& Gao, X.(2019). Insulation Condition Monitoring in Distribution Power Grid via IoT-Based Sensing Network. *IEEE Transactions on Power Delivery*, 34(99):1706-1714.
- [13] Qian, P. , Ma, X. & Zhang, D.(2019). Data-Driven Condition Monitoring Approaches to Improving Power Output of Wind Turbines. *IEEE Transactions on Industrial Electronics*, 66(8):6012-6020.

- [14] Gonzalez-Hernando, F., San-Sebastian, J. & Garcia-Bediaga, A.(2019). Wear-Out Condition Monitoring of IGBT and mosfet Power Modules in Inverter Operation. *IEEE Transactions on Industry Applications*, 55(6):6184-6192.
- [15] Lee, C. Y ., Dong, Z. H. (2019). Hierarchical Equipment Health Index Framework. *IEEE Transactions on Semiconductor Manufacturing*, 32(3):267-276.
- [16] Li, B ., Zhao, R . & Lu, J.(2021). Control on Abnormal Data Overflow of Distribution Network Management Platform. *Journal of Physics Conference Series*, 1748(3):032064.
- [17] Chu, Z. , Wang, W. & Li, B. (2021). An operation health status monitoring algorithm of special transformers based on BIRCH and Gaussian cloud methods. *Energy Reports*, 7(3):253-260.
- [18] Fei, J., Yao, Q .& Chen, M.(2020). The Abnormal Detection for Network Traffic of Power IoT Based on Device Portrait. *Scientific Programming*, 2020(9):1-9.
- [19] Zhao, J ., Mili, L.(2019). A Theoretical Framework of Robust H-Infinity Unscented Kalman Filter and Its Application to Power System Dynamic State Estimation. *IEEE Transactions on Signal Processing*, 67(10):1-12.
- [20] Xiao, G. X., Yan, W. & Li, J. (2021). Simulation of Anti Interference Monitoring of Low Voltage Power Line Leakage Based on ACO. *Computer Simulation*, 38(06):56-60.

# Automated Pneumonia Diagnosis using a 2D Deep Convolutional Neural Network with Chest X-Ray Images

Kamila Kassylkassova<sup>1</sup>, Batyrkhan Omarov<sup>2</sup>, Gulnur Kazbekova<sup>3</sup>, Zhadra Kozhamkulova<sup>4</sup>, Mukhit Maikotov<sup>5</sup>, Zhanar Bidakhmet<sup>6</sup>

L.N.Gumilyov Eurasian National University, Astana, Kazakhstan<sup>1</sup>

Al-Farabi Kazakh National University, Almaty, Kazakhstan<sup>2,6</sup>

International University of Tourism and Hospitality, Turkistan, Kazakhstan<sup>2</sup>

Khoja Akhmet Yassawi International Kazakh-Turkish University, Turkistan, Kazakhstan<sup>3</sup>

Almaty University of Power Engineering and Telecommunications<sup>4,5</sup>

**Abstract**—Tiny air sacs in one or both lungs become inflamed as a result of the lung infection known as pneumonia. In order to provide the best possible treatment plan, pneumonia must be accurately and quickly diagnosed at initial stages. Nowadays, a chest X-ray is regarded as the most effective imaging technique for detecting pneumonia. However, performing chest X-ray analysis may be quite difficult and laborious. For this purpose, in this study we propose deep convolutional neural network (CNN) with 24 hidden layers to identify pneumonia using chest X-ray images. In order to get high accuracy of the proposed deep CNN we applied an image processing method as well as rescaling and data augmentation methods as shear\_range, rotation, zooming, CLAHE, and vertical\_flip. The proposed approach has been evaluated using different evaluation criteria and has demonstrated 97.2%, 97.1%, 97.43%, 96%, 98.8% performance in terms of accuracy, precision, recall, F-score, and AUC-ROC curve. Thus, the applied deep CNN obtain a high level of performance in pneumonia detection. In general, the provided approach is intended to aid radiologists in making an accurate pneumonia diagnosis. Additionally, our suggested models could be helpful in the early detection of other chest-related illnesses such as COVID-19.

**Keywords**—Pneumonia; deep learning; CNN; chest X-rays; radiology

## I. INTRODUCTION

Pneumonia is an infection of the lungs that can be caused by viruses, bacteria, or fungus. In principle, the condition is brought on by a wide range of bacteria, including fungal environmental pollutants, or even physical harm to the lungs caused by smoking or other forms of pollution [1]. Being a common disease, it also has the potential to be fatal [2]. Regarding modern solutions for pneumonia detection, it is hardly surprising that computer vision is an important field of study for artificial intelligence (AI), primarily since it provides answers to a wide range of issues that nowadays people encounter. Biomedical image analysis using artificial intelligence is one of the computer vision areas that has repeatedly shown to be beneficial.

Recent years have seen a rise in the use of deep models, particularly CNN models, as the dominant method for the

categorization of clinical imaging. This is due to the fact that selecting which features to retrieve in conventional approaches to computational intelligence is a time-consuming process that also varies depending on its object [3-5]. These studies have made use of CNNs with a variety of architectural styles and methodological approaches, which were conducted using X-ray images. In order to get more favorable outcomes, CNN-based models require a substantial number of training samples [6]. It is quite problematic to gather medical imaging due to the process of collecting and identifying healthcare data which is complicated by time-consuming privacy rules and the explanations of medical professionals. According to researchers [7], transformation-based data augmentation has been proved to be a suitable method for classifying images. Image enhancement methods may help prevent overfitting during the training phase, which ultimately results in a more accurate model being created.

The majority of the strategies discussed here use transfer learning, which means the deep learning methods were initially trained on the data unrelated to pneumonia diagnosis. The usage of convolutional neural networks created from scratch in a number of image processing algorithms [8] shown that a simpler structure might achieve a higher precision than numerous pre-trained older models used in transfer learning. Throughout this work, we propose creating a cutting-edge deep convolutional neural network as a precise solution for the pneumonia detection issue. The primary innovation in the proposed network is the use of dropout in the convolutional portion of the model, as compared to a large number of other models, which only employ dropout in the fully connected portion of the neural network, which is where the bulk of the attributes are learned. This study establishes the suggested network's ability to perform correct classification even with a small number of training features, providing accompanying illustrations as to how this characteristic might increase predictive performance.

The remainder of this study is as follows: next section reviews the related works in the area of pneumonia detection. Section III describes the data used in this study and the division of the data into training, test and validation sets. Section IV

describes materials and methods including the proposed deep CNN for automated pneumonia detection. Section V demonstrates the obtained results and compares them with the state-of-the-art studies. Section VI discusses the proposed model referencing advantages, challenges and future perspectives. Finally, we conclude the study referencing the main results.

## II. RELATED WORKS

In the past few years, the subject of artificial intelligence known as "deep learning" has had an incredible amount of growth, quickly becoming an integral feature for computational intelligence in a variety of domains, including text processing, speech analysis, image analysis, audio and video processing, and etc. CNNs, have shown themselves to be useful tools for a broad variety of applied problems. The common driving factors of this phenomenon are considered to be an appearance of massive datasets as well as increasingly powerful computing systems. To bridge the gap from high-level representation and low-level attributes, CNNs take as input source data, such as photos, and then execute a series of convolutional operations to gain knowledge rich information about the images. This ensures that the inputs are accurately mapped to the original values. Impressively, it has been recently discovered that CNNs can correspond or even surpass human quality in visual problems such as the classification of image features.

Pneumonia is a leading cause of mortality, accounting for more than four million deaths annually [9]. In terms of timely detection methods, chest X-rays is a prominent diagnostic tool for pneumonia. As for its advantages, chest X-rays are less expensive and can be performed anywhere in the world. Since the indications of pneumonia in X-ray data are not always coherent or readable to the human eye, the software may assist the medical expert in making a diagnosis of pneumonia owing to its fast rate and objective repeatable judgment, not least because of the computer has a high degree of objectivity. Besides, there are a few different decision making systems available, which may be used for the diagnosis and classification of pneumonia. It should be noted that the COVID-19 pandemic was the impetus for a fresh wave of research towards the creation of such diagnostic techniques. In due course, deep learning techniques will provide the backbone of the majority of the innovative approaches. In order to diagnose viral pneumonia [10] used VGG-16 architecture, Resnet-50 convolutional neural network, and Inception-V3 image recognition model, as well as transfer learning, and obtained from 71% to 88% accuracy on around 600 test photos.

Brunese et al. [11] developed a multiclass pneumonia identification system that was based on VGG16 structure and visual debugging method. Using this framework, they were able to obtain an accuracy of 96.2% on a collection of 6500 CXR pictures. When Panwar et al. [12] paired the VGG-19 network and GradCAM strategy, they managed to attain an overall accuracy of 95.6% within the context of a three-class pneumonia classification. In a similar manner, Ibrahim et al. [13] provided a method for the diagnosis of three classes of pneumonia by using VGG and Res-Net152V2 networks. Based

on their own collection of X-ray pictures, they got recall values ranging from 93–97%.

A surprising 98.62% efficiency was achieved using a three-stage hybrid model that was suggested by Jin et al. [14] and consisted of a feature representation, a feature selection, and a SVM classification stage. Karthik et al. [15] used the Channel-Shuffled Dual-Branched convolutional neural network to differentiate between several forms of pneumonia using a variety of datasets that were made accessible to the public. The researchers achieved scores ranging from 94% to 98%. When Quan et al. [16] applied combination of the DenseNet and CapsNet architectures, they obtained a recall of 96% on a COVID-19 dataset with a limited sample size, but with an accuracy of 90.7% on a bigger collection of pneumonia X-ray data. Alhudhaif et al. [17] applied a convolutional neural network that is 201 layers deep and obtained 90% and 95% accuracy and recall values on 6000 X-ray images, using the dataset provided by Kaggle. The design known as Mask R-CNN was used by Jaiswal et al. [18] in order to identify specific areas of the lung that were affected by pneumonia. The so-called CovXNet approach was developed by Mahmud et al. [19], a convolutional neural network that extracts a wide range of attributes from X-ray images by using depthwise convolution with varying dilation convolution rates. Based on 4697 X-ray pictures, Wang et al. [20] trained a visual geometry group architecture using what is likely the most extensive CXR picture dataset that is currently available. This dataset was released by the RSNA Pneumonia Detection Challenge and contains more than 120,000 individual images. When it came to the diagnosis of pneumonia, they had a success rate of 94.62%. The reader is encouraged to consult more recent review studies, such as [21-23], for further information about the current state of the art.

## III. DATA

The dataset consists of 5863 chest X-ray images collected in 2018 from the Women's and Children's Medical Center and the Laboratory of Regenerative Medicine and Healthcare in Guangzhou City [24]. Chest X-ray scans were taken as part of standard clinical examination, and the study objectives didn't include collecting data for this task. The data has been divided into Training, Test, and Validation sets, each of them containing subfolders for the image category as normal and pneumonia. X-ray images have a size of more than 1000 pixels per measurement and occupy more than 1.2 GB of memory.

Table I demonstrates train/test split of the applied dataset. Normal is a label for an X-ray of the healthy chest image, without signs of infectious and inflammatory lung diseases. Pneumonia label is a sign that characterizes a sick patient who has been diagnosed with pneumonia (Fig. 1). Before publication, all chest images were initially checked by removing low-quality unreadable images, as well as by muffling noise in the images. Diagnoses were determined by two highly qualified doctors. In order to prevent human error, all images were retested by a third expert before the machine learning model was trained. All these tests provide confidence in the authenticity and data integrity.

TABLE I. DATASET DESCRIPTION

Dataset	Normal	Pneumonia	Percentage	Dataset
Training	1341	3875	89.68%	Training
Test	234	390	6.7%	Test
Validation	8	8	0.28%	Validation

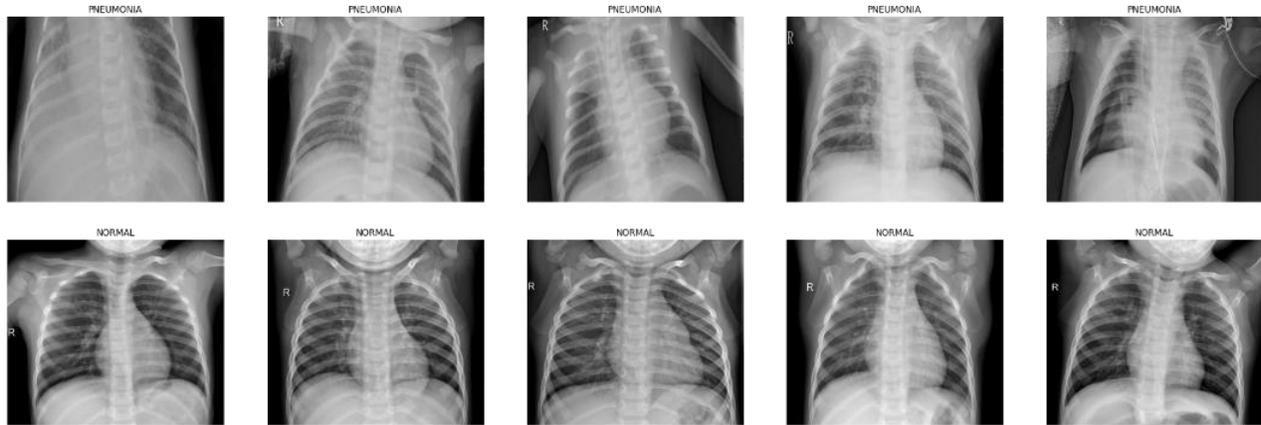


Fig. 1. Chest X-rays of patients

#### IV. MATERIALS AND METHODS

##### A. Block Diagram

A block diagram was built to show in detail what steps you need to go through and how to implement the model. A graphical representation of the entire process using diagrams and blocks is the best way to demonstrate the work process to audience. In its turn, our project has similar development stages, like in another image classification tasks. Each block

shows one development step. Sub blocks may occur within the blocks. They indicate a particular stage for the methods or functions that will be used. Arrows describes the right sequence of working process in our project. In total, creating an algorithm for classifying chest images takes 7 development stages. The process of solving the classification of X-ray images can be represented as a sequence of steps shown in Fig. 2.

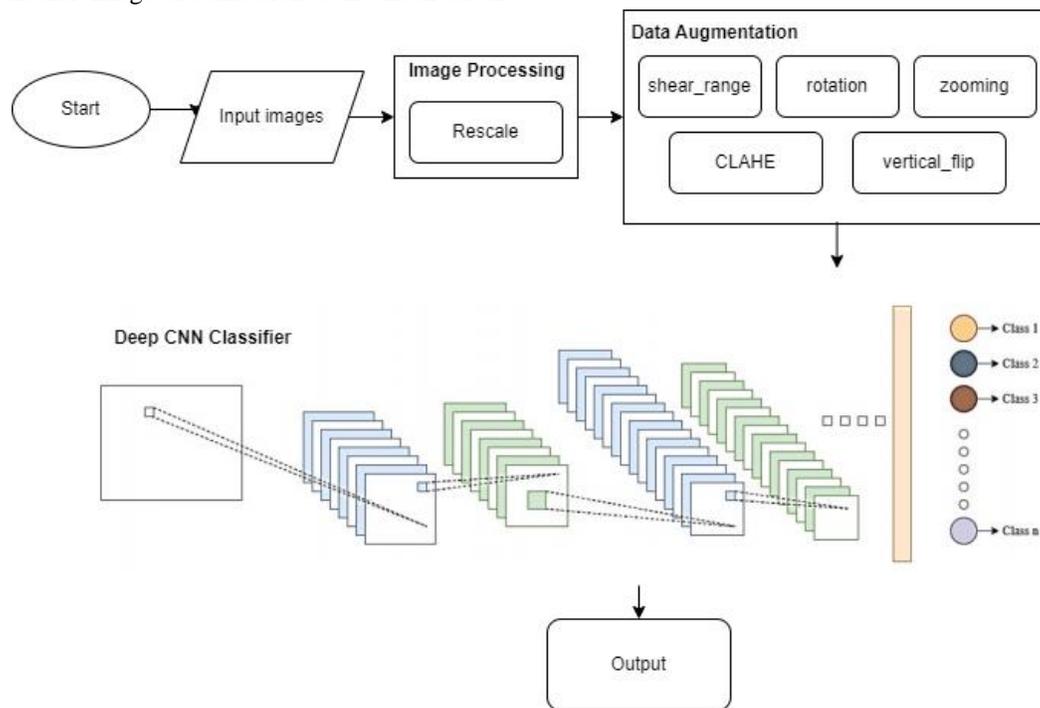


Fig. 2. The proposed framework architecture

The scheme begins with data input, this is the initial part where images are selected for further processing and submission to the model input. The first step of the algorithm is scaling, a constant that will multiply the data before any other processing. Since our source images consist of an RGB color model in the range 0-255, these values are very difficult to calculate. Therefore, we focus on values between 0 and 1, scaling by a factor of  $1/255$ , which is followed by data augmentation. Initially, the data has an imbalance of classes, and data augmentation is a powerful tool that in almost every case helps to increase the efficiency and reliability of the model. Afterwards, the clean data is used to build a CNN architecture. The algorithm takes an input data, assigns importance to various areas/objects in the image, and tries to distinguish one from the other. All parameters are optimized by minimizing the error on the training set by the backpropagation method.

### B. Image Preprocessing

As mentioned earlier, our data has 1000 pixels per chest image and a size of more than 1.2 GB, which means that we would not have enough memory space to simultaneously process such many images. To solve this problem, we needed to find a more efficient way to submit and process data.

To prevent memory loss, a data generator was created using the Python Keras library, which can be employed on top of the GPU installation. It generates our dataset on multiple cores in real-time and immediately transfers to our deep learning model. Separately, the image generator was also built for training, testing, and validation data. For training, we used data augmentation to prevent an imbalance within classes. The rescale argument is defined with the parameter  $1.0 / 255.0$ . After processing the image, we can see the outcome in Fig. 3.

### C. Data Augmentation

In case of class imbalance, Data Augmentation is the main part of image processing, which increases the number of objects of a smaller class as well as the accuracy of the model and creates variability in the data. The general scheme of data augmentation is illustrated in Fig. 4.

As shown in the figure, first the chest images pass through the transformation functions, which are initially determined by a medical specialist. Although there are numerous methods of image processing, such as rotation, shift, brightness, shift inten-

sity, horizontal flip, etc. each individual case employs their own augmentation methods. In our position, the training and test data splitted 80% to 20%. It can be noticed that the classes in the training data are noticeably different, so, in order to solve this problem, we used the following data processing techniques.

A data magnification is a method which generates new images by shifting elements in the image. The methodology is based on a shift or a slice of the image from both the X and Y axes. The measure of the shift or the probability of direction is set by the user as the `shear_range` argument for the `ImageDataGenerator` function. We used a small coefficient so that the change in the shape of the chest is insignificant (`shear_range = 0.2`). Example in Fig. 5.

The technique can change the angles at which our object is located. In the latest updates, images can be changed not only along the horizontal axis, but also to tilt in any direction. Moreover, the rotation method can improve the model and make it more stable on new data. In general, this method is beneficial for overtraining. If all images are served in a fixed position, it can improve the variety of images and increase the variation, which prevents the model from retraining. We used the `rotation_range = 10` argument of the `Image Data Generator` function (see Fig. 6).

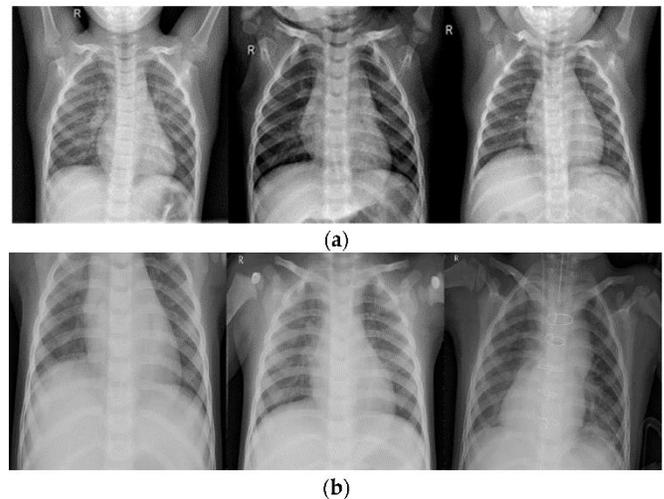


Fig. 3. Image before and after rescaling

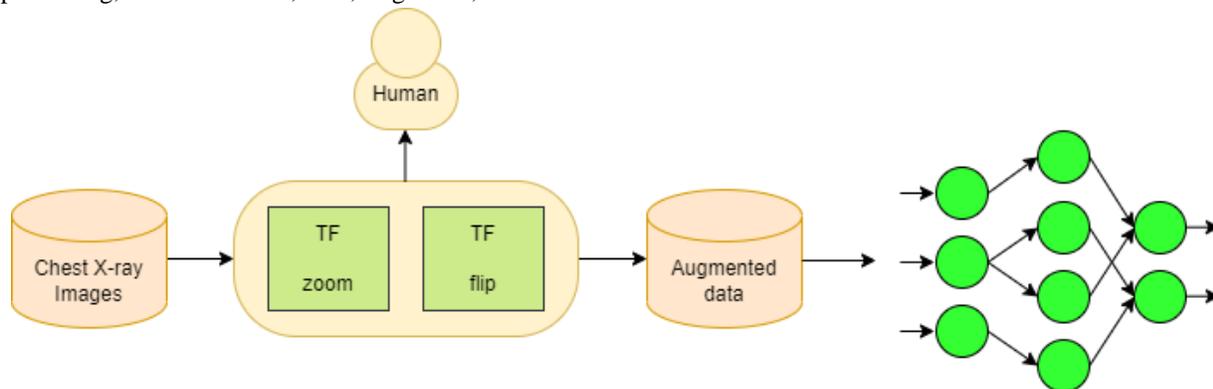


Fig. 4. The scheme of data augmentation (TF - transformation functions)

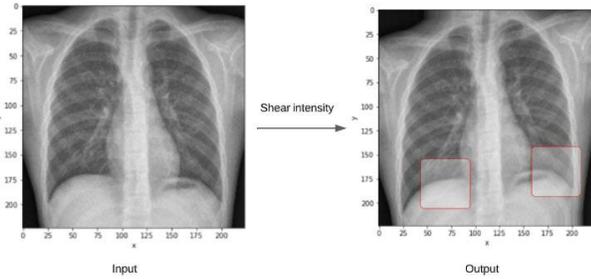


Fig. 5. Image after tuning share intensity



Fig. 6. Rotated image

The method simply increases the scale of the input image and can also add additional pixels around the image. It can take an integer, a floating-point number, and a tuple with numerical values as an argument. There are some restrictions, if the supplied number is less than 1.0, then it enlarges the image, if less than 1.0 it reduces. In our example, a coefficient equal to 0.3 is used.

A processing method that changes the contrast of an image; it is a modified AHE method, it prevents excessive noise amplification in the image, limiting contrast enhancement by setting a threshold that cuts off the histogram before calculating the cumulative distribution function. To use it in our data processing, a function was created, and was passed as an argument to the preprocessing\_function method ImageDataGenerator.

#### D. Proposed Model

Today convolutional neural networks are extensively used for classification, image segmentation and object detection projects. The principle of the algorithm is as follows, it assigns importance to each area of the image, and based on the weights, learn to distinguish one from the other. The architecture of a convolutional neural network is illustrated in Fig. 7.

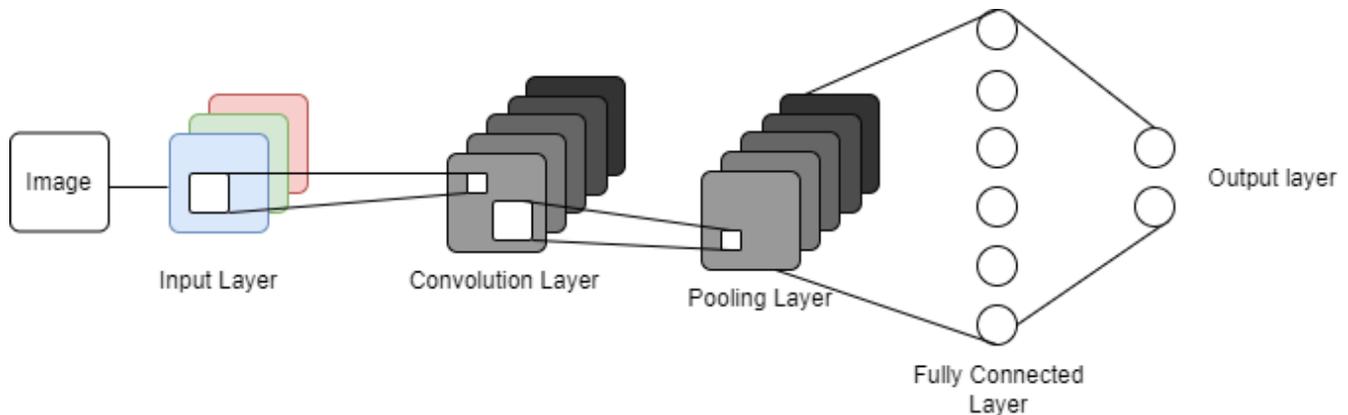


Fig. 7. Convolutional neural network architecture

As shown in the picture, the training steps can be divided into several stages: input layer, convolution layer, pooling layer, fully connected layer and output layer.

The first step of the algorithm is to prepare the image in the correct format to pass the model. In order to achieve this goal, we translated the images into a matrix view. If the input images have a black-and-white appearance, then in matrix form they will be of size  $n = m$ , which means the matrix will be in two-dimensional. Nevertheless, most often in the real world, the data will be colored, and due to this they will need to be stored in a three-dimensional matrix. In addition to width and height, the data includes a third channel, which is called chromatic. As soon as the images are ready, we move to the convolution layer.

Convolutional layer is considered as the main CNN layer. At this stage, the matrix form of our image takes a filter (core). The number of filters is determined arbitrarily with a size of  $3 \times 3$  or  $5 \times 5$ . This filter passes through the image and extract activation cards. During the pass, the filter is multiplied by the pixels in the image and summed by the formula:

$$x_j^l = f\left(\sum_i x_i^{l-1} * k_j^l + b_j^l\right) \quad (1)$$

Here,  $x_j^l$  is a feature map I (output of l layer),  $f(x)$  is an activation function, "\*" is a convolution operation of input x of the map, layer l,  $b_j^l$  is the coefficient of the layer l for the feature map j.

Pooling layer – the main goal in a CNN is to collect and reduce the dimension of the activation map. Regardless of what kind of unifying layer you use, it will always reduce the amount of computing effort of the network. The main two functions that are most often used in practice are:

- Average Pooling: Calculates the average value for its plot.
- Max Pooling: Selects the maximum value of the field in the activation map.

A fully connected layer is the most recent and basic layer of a convolutional neural network. It provides our ready and processed data to identify the answer. A fully connected layer combines all the previous layers, smoothest and transforms into a vector view.

In the last fully connected layer, various activation functions are used. As an activation layer, we take rectified linear unit (ReLU).

Our developed model consists of six convolutional neural networks. For the first layer, preliminarily processed images are submitted in a matrix format with a size of  $224 \times 224 \times 3$ . After each layer, max pooling is applied to lower the activation card. In the last block, convolution alignment is applied and sent as an input layer for the fully connected layer. The last output of the fully connected layer is two nodes that show whether a person is sick or not. To determine it, the softmax activation function is used. Adam was chosen as the optimization algorithm for the convolutional neural network. Thanks to this method, weights and learning rates are selected and changed, as well as model losses are reduced.

In order to characterize the correctness of our solutions, the loss function categorical cross entropy is used. The loss function has the following general form:

$$x_j^l = f\left(\sum_i x_i^{l-1} * k_j^l + b_j^l\right) \quad (2)$$

Here:  $\hat{y}_i$  is the  $i$ -th response of the model;

$y_i$  is the corresponding true value;

*outputsize* is the number of scalar values in the output of the model;

This loss describes our model very well and shows how two discrete probability distributions are distinguishable from each other. Architecture of the entire convolutional neural network is illustrated in Fig. 8.

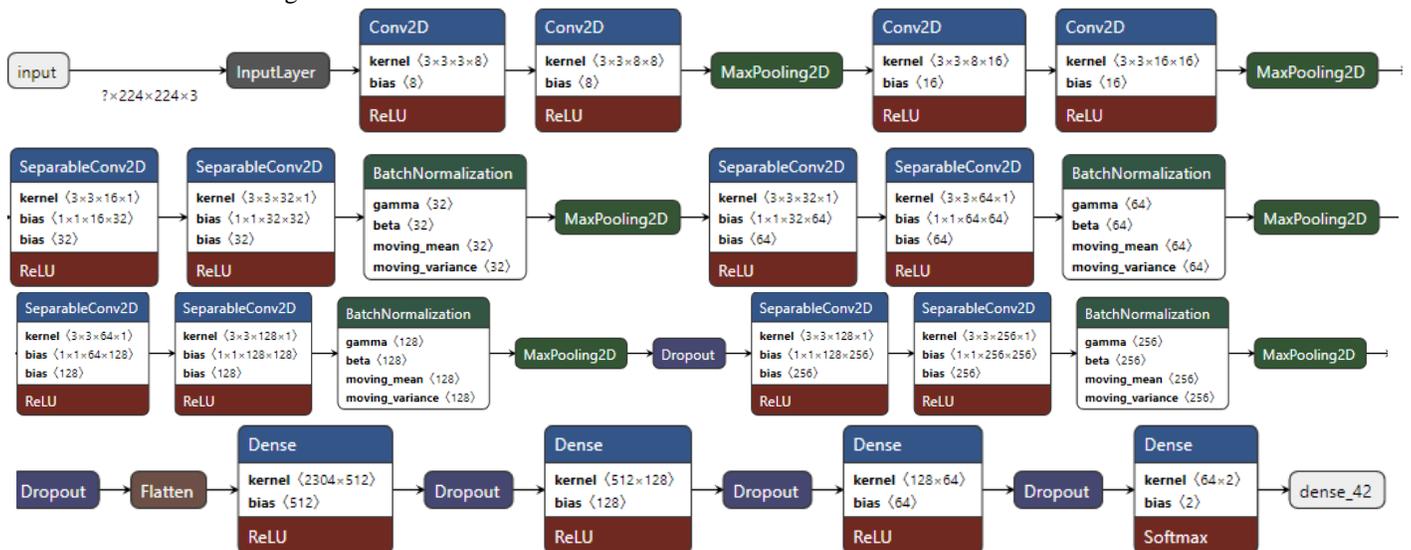


Fig. 8. Architecture of the proposed CNN model

### E. Normalization

Normalization is one of the main methods of preprocessing. It is used to standardize data, in short, converts data into the same range to reduce the computational time of the algorithm. The lack of normalization can lead to the complexity of the net-work and reduce the learning rate. In addition to standardization, it also helps to regularize the CNN model, which is one of its unintended advantages.

Specifically for our model, we used batch normalization. It is important not to get confused, since batch normalization is not used for source data, and it is activated between network layers. The data is divided into small chunks with an average value of 0 and a standard deviation of 1. After processing, each element of the original data simulates a standard normal distribution. Here you can see the batch normalization formula.

$$Z^N = \left( \frac{z - m_z}{s_z} \right) \quad (3)$$

Where:  $s_z$  – standard deviation of output data (neurons);

$m_z$  – average value of output data (neurons);

$z$  – input value for normalization;

### F. Evaluation Metrics

Evaluation criteria include Accuracy, Precision, Recall, and AUC-ROC curve. In this section, we explain each of the evaluation parameters that applied in current research.

Confusion Matrix: With the help of the confusion matrix, we can see in more detail how our model behaves in various situations (see Fig. 9). It not only shows the model's response errors, but also their types. Thus, such breakdown of the response helps to prevent limitations related to accuracy.

Accuracy: This is the ratio of the proportion of correct answers of the model to all its variables. For binary classification, the precision formula is defined as

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (4)$$

Precision: This is the ratio of the model's true-positive responses to the total number of positive predictions. For example, if the precision value turned out to be 0.6, it means that if the model predicts that the patient has pneumonia, then it is correct for 60% of cases. Precision formula:

$$Precision = \frac{TP}{TP + FP} \quad (5)$$

Recall: The recall characterizes the ability of the model to identify correctly positive samples. This means that for a good classification model, the recall should be closer to one. For example, if recall = 0.8, it means that the model determines 80% of pneumonia diseases. The formula for recall:

$$Recall = \frac{TP}{TP + FN} \quad (6)$$

AUC-ROC: The Receiver Operator Characteristic (ROC) is a probability curve that shows the ratio of TPR and FPR. They, in turn, separate the "signal" from the "noise". The Area under the Curve (AUC) is the area under this curve. The wider and larger it is, the better the classification model works.

### V. EXPERIMENTAL RESULTS

The accuracy of the model for training data has reached up to 91%, which indicates that the model classifies chest images adequately. However, it would be unpractical to reckon only one metric, in the light of an imbalance of classes. Therefore,

we compared other metrics as well, and Precision and recall are kept around 90 percent, and the quality of validation data has decreased. This happens because on the validation data we have only 16 new images that the model has not seen before, so the system operated correctly on 13, only 3 images have an erroneous result, which is a very good indicator.

More detailed statistics can be seen in Fig. 10, where the model's responses are recorded point-by-point for each case. Our neural network correctly found 372 patients with pneumonia and identified 194 healthy patients, while in some images the system mistakenly classified patients. In most cases, there is an ambiguity in the pictures, or the low quality of the pictures.

Fig. 11 demonstrates the AUC-ROC curves obtained by the proposed deep CNN on all the five folds of cross-validation. The obtained results show that, the proposed deep CNN gives high accuracy between 0.979 to 0.988 AUC-ROC value in pneumonia detection problem.

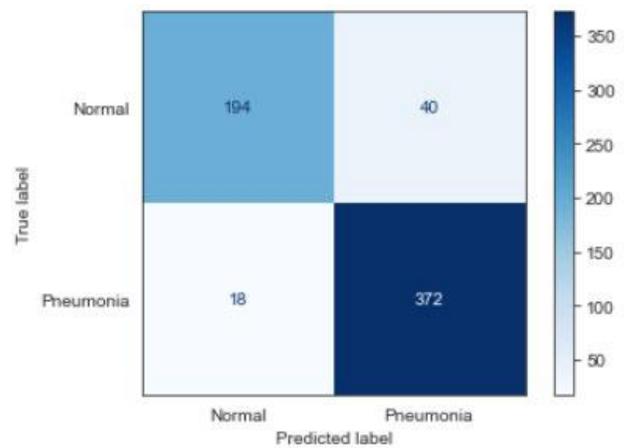


Fig. 9. Confusion matrix.

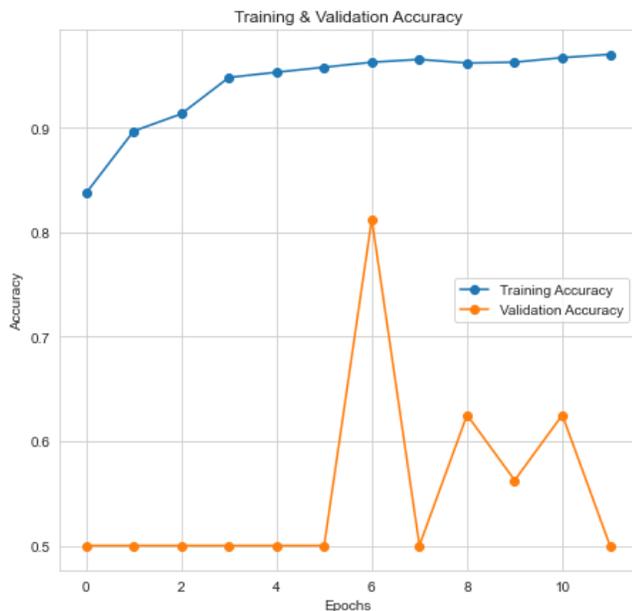
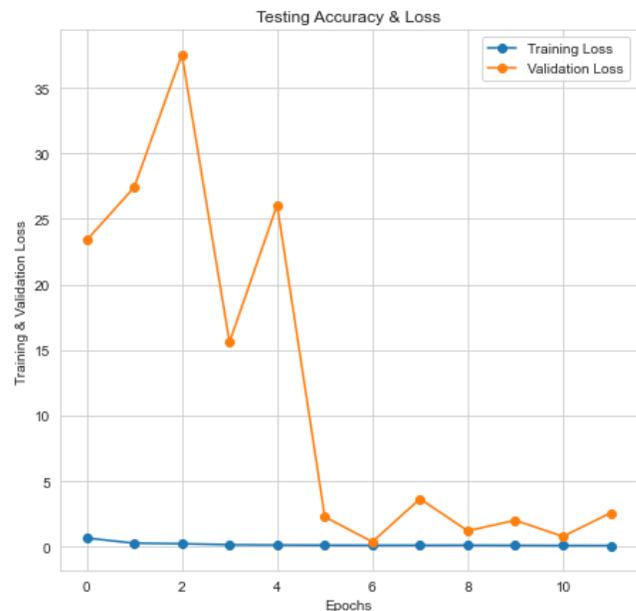


Fig. 10. Training and validation accuracy



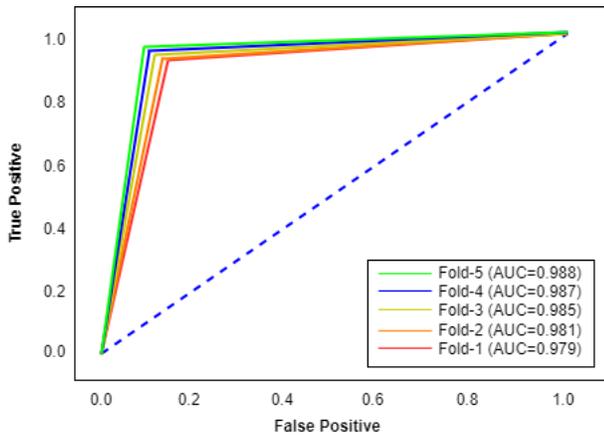


Fig. 11. AUC-ROC curve of the obtained results.

Fig. 12 demonstrates several examples of the obtained results using the proposed deep convolutional neural network, while Fig. 12(a) illustrates the case that the proposed model found no trace of pneumonia when ground truth coincides with that decision. Fig. 12(b) demonstrates that there is a pneumonia when the ground truth is pneumonia. Fig. 12(c) shows the case that the model found as a pneumonia, when there is no pneumonia that coincides the false positive case.

Table II compares the proposed deep convolutional neural network with the state-of-the-art studies that dedicated to deep learning based pneumonia detection. The results show that the proposed deep convolutional neural network shows high performance in terms of different evaluation parameters including accuracy (97.2%), precision (97.1%), recall (97.43%), F-score (96%), and AUC (98.8%).



a) Predicted class is false, ground truth is false    b) Predicted class is true, ground truth is true    c) Predicted class is true, ground truth is false

Fig. 12. Experimental results.

TABLE II. DATASET DESCRIPTION

Approach	Dataset	Number of images	Results
Proposed model	Kaggle CXR dataset [24]	5863	97.2% accuracy, 97.1% precision, 97.43% recall, 96% F-score, 98.8% AUC
EfficientNet B3 Model [25]	Dataset of X-ray images	13569	93.9% accuracy, 96.8% recall
COVID-net model [26]	COVIDx dataset	13975	92.4% accuracy, 88.6% recall
COVID-ResNet model [27]	Kaggle's COVID19 Global Forecasting dataset [28]	13945	96.23% accuracy, 96.92% recall
Customized CNN [29]	Chest X-Ray dataset [30]	5856	93.6% accuracy
CNN [31]	Chest X-Ray dataset [30]	5840	94.40% accuracy
VGG 16 [32]	Chest X-Ray dataset [30]	5856	84.5% accuracy
DenseNet-121 [33]	PadChest dataset [34]	5232	92.8% accuracy

## VI. DISCUSSION AND FUTURE RESEARCH

The experiments' results show that the proposed model, which is a lightweight CNN, not only performs a smaller number of calculations compared to the majority of deep learning models, but it also performs better in terms of important evaluation parameters as accuracy and recall. This can be observed when the number of variables is almost equal to an order of magnitude. Therefore, utilizing the depthwise separable convolution is beneficial in this situation. In light of the advent of deep learning, the majority of image processing

models now includes a huge number of attributes and need a significant amount of computation, making them unsuitable for implementation in embedded devices. When it comes to diagnosing pneumonia, which is a fairly frequent condition, there is also necessity to consider how to complete it in a timely and accurate manner in locations where medical equipment and medical professionals are unavailable. One of the reasons why we suggest utilizing the proposed deep CNN for pneumonia diagnosis is because of this particular benefit.

Recently some studies [15, 16] tried to develop a model from scratch or to adapt an existing model for the aim of identifying pneumonia. Next researches [13,14,17] concentrated on applying transfer learning and pretrained models. Meanwhile, Mahmud et al. [19] utilized a dataset from Kaggle [34], while Wang et al. [20] used training data from Mendeley, while the majority of studies [29-32] applied the Chest X-ray dataset [30]. In contrast to this work, a few other scholars [13, 15] did not employ any kind of data augmentation technique in their work.

Table II demonstrates that the proposed model has a recall of 97.43%, which makes sensitivity the most essential parameter in medical applications since it reveals the proportion of positive diagnoses that are accurate. When compared to the approach that has been proposed, all of the prior work [25-27, 29, 31-36] has been shown to produce a lower accuracy, sensitivity, precision, and F1 score than the way that has been recommended. In addition to this, the size of the dataset that they employed is far lower than the one that was used in this work. The structure of the suggested model is straightforward, which indicates that it converges quickly and does not call for a significant amount of processing resources. On the other hand, the generalization power of the recommended model is not on par with that of pretrained models. Only 5852 CXR images were utilized for the training of the suggested architecture, in contrast to the millions of photos that were used for the training of the pretrained models. The picture dataset that was employed for this research is insufficient to develop a credible CNN model with a high level of accuracy and to include all of the intrinsic image characteristics associated with pneumonia. In conclusion, the researchers who carried out the earlier studies did not provide sufficient information to carry out an exhaustive analysis; furthermore, they did not divulge the methodology that they applied in order to validate their data, and it is unclear whether or not they made use of cross validation like the current investigation did.

The fact that this study came up with almost faultless outcomes lends credibility and dependability to the applied methodology [37-40]. Lastly, the recommended model will have an influence on medicine. With the help of the proposed models, medical professionals working in rural regions will be able to diagnose pediatric pneumonia in a timely manner that is both cost-effective and accurate. In terms of older patients and younger children in particular, prompt and precise diagnosis of pneumonia may lessen the likelihood of deadly consequences from the disease. The interpretation variability and subjectivity issue that might arise while reading a chest X-ray radiograph can be helped by the suggested model. It may also be used to help inexperienced radiologists in distant places that lack professional radiologists to make a proper choice about a patient's treatment. The last phase is to create a mobile app that can differentiate between pneumonia and chest X-ray images, which will be used by airport personnel.

## VII. CONCLUSION

Our final task is to carry out this project in medical institutions, since AI helps hospitals, especially those with limited resources, quickly to examine suspected patients for

further diagnosis and treatment. In a few seconds after taking chest-X ray, the radiologist receives a notification about whether the patient needs to be assigned a high priority and enter a protocol for the treatment of pneumonia or not. More traditional methods may take longer to process due to an increase in the number of infected. Thus, AI has three valuable features during a serious outbreak:

- 1) Patients with symptoms are admitted to hospitals in large numbers, while AI can help to prioritize patients quickly.
- 2) AI serves as a supplement to the diagnosis of pneumonia, since the capabilities of the laboratory, even in cities with good equipment, are insufficient to work at a pace with the increase in the number of suspected cases. AI is an important supplement in an outbreak of a disease with high infectivity.
- 3) AI can easily compare changes and events in the lungs with different examinations of the same patient, which can be tedious and difficult for doctors, especially in the context of an epidemic situation.

All things considered, we firmly believe that that our project helps doctors for diagnosing pneumonia, provide immediate help to sick people and save time for doctors as well as patients.

## REFERENCES

- [1] Saragih, P., Sihombing, V. E., & Pardede, I. B. Y. (2022). Factors that cause the increase of pneumonia in Indonesia. *AMCA Journal of Community Development*, 2(1), 31-33.
- [2] Zou, X., Suo, L., Wang, Y., Cao, H., Mu, S., Wu, C., ... & Cao, B. (2022). Concurrent pi-geon paramyxovirus-1 and *Acinetobacter baumannii* infection in a fatal case of pneumonia. *Emerging Microbes & Infections*, 11(1), 968-977.
- [3] B. Omarov, N. Saparkhojayev, S. Shekerbekova, O. Akhmetova, M. Sakypbekova et al., "Artificial intelligence in medicine: real time electronic stethoscope for heart diseases detection," *Computers, Materials & Continua*, vol. 70, no.2, pp. 2815-2833, 2022.
- [4] Shukayev, D. N., Kim, E. R., Shukayev, M. D., & Kozhamkulova, Z. (2011, July). Modeling allocation of parallel flows with general resource. In *Proceeding of the 22nd IASTED International Conference Modeling and simulation (MS 2011)*, Calgary, Alberta, Canada (pp. 110-117).
- [5] Moshkalov, A. K., Iskakova, M. T., Maikotov, M. N., Kozhamkulova, Z. Z., Ubniyazova, S. A., Stangazyeva, Z. K., ... & Darkhanbaeyeva, G. S. (2014). Ways to improve the information culture of students. *Life Science Journal*, 11(8s), 340-343.
- [6] B. Omarov, A. Tursynova, O. Postolache, K. Gamry, A. Batyrbekov et al., "Modified unet model for brain stroke lesion segmentation on computed tomography images," *Computers, Materials & Continua*, vol. 71, no.3, pp. 4701-4717, 2022.
- [7] Erdem, E., & AYDİN, T. (2021). Detection of Pneumonia with a Novel CNN-based Ap-proach. *Sakarya University Journal of Computer and Information Sciences*, 4(1), 26-34.
- [8] Kumarasinghe, H., Kolonne, S., Fernando, C., & Meedeniya, D. (2022). U-Net Based Chest X-ray Segmentation with Ensemble Classification for Covid-19 and Pneumonia. *International Journal of Online and Biomedical Engineering (iJOE)*, 18(07), pp. 161-175.
- [9] Mujahid, M., Rustam, F., Álvarez, R., Luis Vidal Mazón, J., Díez, I. D. L. T., & Ashraf, I. (2022). Pneumonia Classification from X-ray Images with Inception-V3 and Convolutional Neural Network. *Diagnostics*, 12(5), 1280.
- [10] Jayakumar, L., Chitra, R. J., Sivasankari, J., Vidhya, S., Alimzhanova, L., Kazbekova, G., ... & Teressa, D. M. (2022). QoS Analysis for Cloud-

- Based IoT Data Using Multicriteria-Based Optimization Approach. Computational Intelligence & Neuroscience.
- [11] L. Brunese, F. Mercaldo, A. Reginelli, and A. Santone, "Explainable deep learning for pulmonary disease and coronavirus COVID-19 detection from X-rays," *Computer Methods and Programs in Biomedicine*, vol. 196 p. 105608 (2020)
- [12] Panwar, H., Gupta, P. K., Siddiqui, M. K., Morales-Menendez, R., Bhardwaj, P., & Singh, V. (2020). A deep learning and grad-CAM based color visualization approach for fast detection of COVID-19 cases using chest X-ray and CT-Scan images. *Chaos, Solitons & Fractals*, 140, 110190.
- [13] Ibrahim, D. M., Elshennawy, N. M., & Sarhan, A. M. (2021). Deepchest: Multi-classification deep learning model for diagnosing COVID-19, pneumonia, and lung cancer chest diseases. *Computers in biology and medicine*, 132, 104348.
- [14] Jin, W., Dong, S., Dong, C., & Ye, X. (2021). Hybrid ensemble model for differential diagnosis between COVID-19 and common viral pneumonia by chest X-ray radiograph. *Computers in biology and medicine*, 131, 104252.
- [15] Karthik, R., Menaka, R., & Hariharan, M. (2021). Learning distinctive filters for COVID-19 detection from chest X-ray using shuffled residual CNN. *Applied Soft Computing*, 99, 106744.
- [16] Quan, H., Xu, X., Zheng, T., Li, Z., Zhao, M., & Cui, X. (2021). DenseCapsNet: Detection of COVID-19 from X-ray images using a capsule neural network. *Computers in biology and medicine*, 133, 104399.
- [17] Alhudaif, A., Polat, K., & Karaman, O. (2021). Determination of COVID-19 pneumonia based on generalized convolutional neural network model from chest X-ray images. *Expert Systems with Applications*, 180, 115141.
- [18] Jaiswal, A. K., Tiwari, P., Kumar, S., Gupta, D., Khanna, A., & Rodrigues, J. J. (2019). Identifying pneumonia in chest X-rays: A deep learning approach. *Measurement*, 145, 511-518.
- [19] Mahmud, T., Rahman, M. A., & Fattah, S. A. (2020). CovXNet: A multi-dilation convolutional neural network for automatic COVID-19 and other pneumonia detection from chest X-ray images with transferable multi-receptive feature optimization. *Computers in biology and medicine*, 122, 103869.
- [20] Wang, J., Bao, Y., Wen, Y., Lu, H., Luo, H., Xiang, Y., ... & Qian, D. (2020). Prior-attention residual learning for more discriminative COVID-19 screening in CT images. *IEEE Transactions on Medical Imaging*, 39(8), 2572-2583.
- [21] Yang, Y., & Mei, G. (2022). Pneumonia Recognition by Deep Learning: A Comparative Investigation. *Applied Sciences*, 12(9), 4334.
- [22] Ali, L. R. . ., Jebur, S. A. ., Jahefer, M. M. ., & Shaker, B. N. (2022). Employing Transfer Learning for Diagnosing COVID-19 Disease. *International Journal of Online and Biomedical Engineering (iJOE)*, 18(15), pp. 31-42.
- [23] Alqudah, Y., Sababha, B., Qaralleh, E., & Yousseff, T. (2021). Machine Learning to Classify Driving Events Using Mobile Phone Sensors Data. *International Journal of Interactive Mobile Technologies (IJIM)*, 15(02), pp. 124-136.
- [24] Pneumonia Detection using Chest XRay dataset. [Online]. Available: <https://www.kaggle.com/code/dhvananrangrej/pneumonia-detection-using-chest-xray>
- [25] E. J. S. Luz, P. L. Silva, R. Silva, L. Silva, G. Moreira et al., "Towards an effective and efficient deep learning model for COVID-19 patterns detection in X-ray images," *arXiv Prepr. arXiv: 2004.05717*, pp. 1-10, 2020
- [26] L. Wang and A. Wong, "Covid-net: A tailored deep convolutional neural network design for detection of COVID-19 cases from chest radiography images," *Scientific Reports*, vol. 10, no. 1, pp. 1-12, 2020.
- [27] D. Dansana, R. Kumar, J. D. Adhikari, M. Mohapatra, R. Sharma et al., "Global forecasting confirmed and fatal cases of COVID-19 outbreak using autoregressive integrated moving average model," *Frontiers in Public Health*, vol. 8, p. 580327, 2020
- [28] <https://www.kaggle.com/c/covid19-global-forecasting-week-2>
- [29] Rajaraman, S., Candemir, S., Kim, I., Thoma, G., & Antani, S. (2018). Visualization and interpretation of convolutional neural network predictions in detecting pneumonia in pediatric chest radiographs. *Applied Sciences*, 8(10), 1715.
- [30] Kermany, D., Zhang, K., and Goldbaum, M. (2018). Labeled optical coherence tomography (oct) and chest x-ray images for classification. *Structural Equation Modeling: A Multidisciplinary Journal*.
- [31] A. A. Saraiva, D. Santos, N. J. C. Costa et al., "Models of learning to classify X-ray images for the detection of pneumonia using neural networks," in *Proceedings of the 12th International Joint Conference on Biomedical Engineering Systems and Technologies*, pp. 76-83, Prague, Czech Republic, 2019.
- [32] Ayan, E., & Ünver, H. M. (2019, April). Diagnosis of pneumonia from chest X-ray images using deep learning. In *2019 Scientific Meeting on Electrical-Electronics & Biomedical Engineering and Computer Science (EBBT)* (pp. 1-5). Ieee.
- [33] Cohen, J. P., Bertin, P., & Frappier, V. (2019). Chester: A web delivered locally computed chest x-ray disease prediction system. *arXiv preprint arXiv:1901.11210*.
- [34] Bustos, A., Pertusa, A., Salinas, J. M., & de la Iglesia-Vayá, M. (2020). Padchest: A large chest x-ray image dataset with multi-label annotated reports. *Medical image analysis*, 66, 101797.
- [35] Anand, M., Sahay, K. B., Ahmed, M. A., Sultan, D., Chandan, R. R., & Singh, B. (2023). Deep learning and natural language processing in computation for offensive language detection in online social networks by feature selection and ensemble classification techniques. *Theoretical Computer Science*, 943, 203-218.
- [36] D. Sultan, B. Omarov, Z. Kozhamkulova, G. Kazbekova, L. Alimzhanova et al., "A review of machine learning techniques in cyberbullying detection," *Computers, Materials & Continua*, vol. 74, no.3, pp. 5625-5640, 2023.
- [37] Doskarayev, B., & Kulbayev, A. (2017). Sport as an important factor of strengthening tolerance (The case of Kazakhstan). *Revista ESPACIOS*, 38(46).
- [38] D. Sultan, A. Toktarova, A. Zhumadillayeva, S. Aldeshov, S. Mussiraliyeva et al., "Cyberbullying-related hate speech detection using shallow-to-deep learning," *Computers, Materials & Continua*, vol. 74, no.1, pp. 2115-2131, 2023.
- [39] Issayev, A., Ortayev, B., Issayev, G., Baurzhan, D., & Gulzhaina, A. (2022). Improving the Supervisory Competence of Future Teacher Trainers with the Help of Innovative Technologies. *World Journal on Educational Technology: Current Issues*, 14(3), 692-703.
- [40] Omarov, B., Orazbaev, E., Baimukhanbetov, B., Abusseitov, B., Khudiyarov, G., & Anarbayev, A. (2017). Test battery for comprehensive control in the training system of highly Skilled Wrestlers of Kazakhstan on National wrestling "Kazaksha Kuresi". *Man In India*, 97(11), 453-462.

# Classification of Human Sperms using ResNet-50 Deep Neural Network

Ahmad Abdelaziz Mashaal<sup>1</sup>, Mohamed A. A. Eldosoky<sup>2</sup>, Lamia Nabil Mahdy<sup>3</sup>, Kadry Ali Ezzat<sup>4</sup>  
Biomedical Engineering Department-Faculty of Engineering, Helwan University, Cairo, Egypt<sup>1,2</sup>  
Biomedical Engineering Department, Higher Technological Institute, 10<sup>th</sup> of Ramadan City, Egypt<sup>3,4</sup>

**Abstract**—Infertility is a disease which scientists around the world are concerned with. The disease of infertility also is a worldwide health concern of many people in the community. The andrologists are continually searching for further developed techniques for any related problems. The intracytoplasmic sperm injection (ICSI) method is a widely recognized strategy for accomplishing pregnancy and considered as one of the best methods for infertility treatment worldwide. Choosing the best sperms are done using the vision through the specimen which is reliant on the abilities and the cleverness of the embryologists and as such inclined to human errors. Subsequently, a system that detects the normal sperms automatically is required for speedy and more precise outcomes. Deep learning approaches are usually effective for classification and detection purposes. This paper uses the Residential Energy Services Network (ResNet-50) deep learning architecture to recognize human sperms after classification of human sperm heads. The ResNet-50 proposed model achieved an accuracy of 96.66%. This proposed model demonstrated its efficiency at the detection of healthy sperms. The healthy sperms are used for the injection into eggs by the andrologists who always look for easier and more advanced methods in order to increase the success rate of ICSI process.

**Keywords**—Healthy sperms; sperm heads; infertility; classification; convolution; ResNet-50

## I. INTRODUCTION

Infertility is the failure of couples to accomplish pregnancy after twelve months despite having sex regularly [1]. The issue of the most couples around all countries is the infertility, around 30-40% of them associated with irregularities of male parameters as reported by WHO [2]. The main mission in the ICSI technique is the determination of the optimum sperms which can be used for fertilization of eggs. Embryologists select healthy sperms relying on the sperm shape by visual evaluation. This operation is difficult, tiring and prone to human errors. So, many researchers tend to AI approaches for finding better solutions for the detection of healthy sperms automatically.

The motivation of the paper comes from the need of laboratory technicians for automatic program to perform selection for healthy sperms that can be injected into oocytes. There are many precautions and conditions for the examination of sperm cells. The oversight processes about principles of quality, observation, testing, handling, protection, and storing of sperms are usually based on WHO guidelines [2]. Additionally, conditions in the space of assessment, capturing and the specifications of an equipment that influence

the vision of semen sample by the laboratory technician for the assessment. There is an extra factor influencing the precision in visual assessment which is the cleverness of the andrologists regarding other parameters that affect the sperm selection. To decrease the time of examination of sperms, computer vision and the processing of images have been utilized. Some previous studies did not prove their efficiency because of staining factors, fundamentally it is useful for the research purposes with neural network architectures that are designed using reinforcement learning [18], [19]. Also, there are many algorithms that are used for this purpose and will be discussed in detail next in the literature review section.

The examination of semen sample mostly performed according to the morphology [24]. Some consultations from the andrologists are needed when the examination of morphology is completed successfully. Subsequently, the right decision is taken about the healthy sperms that are suitable for injection process. According to the morphology also, computer-aided sperm analysis is done using the artificial intelligence [20].

The morphology of sperms is the most important parameter for the assessment of the sperm which defines the range of integrity of the shape and size of sperms. There are many factors that affect the quality of human semen [23]. The morphology of sperms is assessed under microscope after the general test of semen which also indicate the percent of morphology abnormalities [21], [22].

New advancements in optics and robotized microscopy have made experts ready to view and record huge number of images. The WHO guidelines are important for the selection of sperms and for the male infertility recognition. It is essentially to inspect the sperms for checking their validity for the injection [25]. The analysis of sperms is usually done according to the morphology and the motility of sperms.

The assisted reproductive technology (ART) is needed for the recognition of the best ways for the finding the healthy sperms for raising the rates of fertilization, and live birth. This system proved its advantages for getting good results with better accuracy for obtaining the normal sperms, which are referring to the normal sperm heads, which can be utilized in ICSI process after the detection using this system. The healthy sperms are used for the injection into eggs by the andrologists and then the fertilized egg is transferred into the uterus. One of the advantages of this study is its ability to deal with non-stained images that is similar to the captured images from microscope. The studies that deal with stained images are not

applicable because the stain is harmful for the human sperms and make them not suitable for the injection process.

Examination of images from in-vitro tests are normally restricted for andrologists at labs and andrology centers where they are responsible for the sperm processing with special tools. Besides, their perceptions follow particular precautions and rules and the semen microbiome has an impact on the infertility and affect the function of sperms [3]. Currently, the manual selection for healthy sperms can cause problems and it cannot be as accurate as automatic systems for sperm selection. On the other side, the automatic systems are labour saver, and it is not necessary for laboratories to provide skilled technicians for sperm selection in the intracytoplasmic sperm injection process. These reasons motivate us to propose a method which provides automated detection for the best sperms that can be used for increasing the success rate of ICSI process and for getting higher pregnancy success rate.

The main parts of the paper are arranged as follows: Section II explains the literature review for this paper. Section III explains the material and methods, showing the dataset, and describes the stages of preprocessing used in this paper. Section IV describes the proposed model for healthy sperm detection. Section V shows the results of this model. Section VI presents the conclusions of this research and refers to the fields of future work.

## II. LITERATURE REVIEW

Many studies about the analysis of sperms automatically using deep learning and machine learning that is important for the process of the intracytoplasmic sperm injection. Jason et al. presented a model for sperm classification of sperms by VGG16 architecture with high accuracy [4]. Soroush et al. introduced a model for the assessment of sperms with three parts [5]. Abolghasem et al. presented a method for sperm recognition with tail and neck with accuracy of 93.2% [6]. Miahi et al. improved an architecture for vacuole abnormality recognition with an accuracy of 91.66% [7]. Prabakaran et al. presented a CNN network for the detection of abnormal sperms with abnormal dimensions with 98.99% of accuracy and approved the efficiency of the program [8]. P., Zuhdi et al. [9] introduced DeepSperm, a profound brain network that utilizes a particular discovery layer to identify little items. The creators expanded the info goal of the organization, utilized a dropout layer and played out an information increase procedure in view of immersion and openness changes. The acquired outcomes beat the cutting-edge regarding accuracy, speed and computational calculations. S. Hicks et al. proved the importance of deep learning in the processing of videos of movable sperms. They used three videos in which labelled manually by the andrologists, the labels were bounding boxes for the egg, catheter and sperm. YOLOv5 model was used for tracking and detection of sperms. The three used videos were separated as two videos for training and one video for validation. The accuracy of eggs' tracking reaches 92% and for ICSI pipette reaches 94%. The author concludes the importance of deep learning for the detection of sperms for performing successful intracytoplasmic sperm injection process [10].

Ruth Marín et al. proposed the deep learning for and studied the effectiveness of transfer learning for the segmentation of sperms. They used public dataset (SCIAN-SpermSegGS) in which evaluated using Mask RCNN and U-net. The U-net model reaches 95% of segmentation for sperm head which consider a promising result for using the computer assisted systems for this purpose [11]. Lee et al. developed an algorithm for the detection of the human sperms with machine learning using microsurgical testicular sperm extraction utilizing bright-field microscopy in which the sperms were collected from healthy men. The algorithm used the CNN based on the U-Net architecture. In this study, the algorithm achieved 91% PPV and this make the rare sperms can be detected with biopsy samples using machine learning [12].

Raffael Golomngi et al. presented automated deep convolutional network the deep learning VGG19 was used for the implementation with accuracy reaches 93%. This method used labelling with bounding boxes for the detection of sperms. This study shows its success on images acquired by the optical microscope [13]. Aristoteles et al. used deep learning approaches for healthy sperms detection using video tracking using deep learning with result 90.31% average precision sperms and the system detects the non-sperm parts automatically rather than the manual methods that require more time and high costs [14]. Ahmad Mashaal et al. proposed a system for the recognition of healthy sperms using VGG16 deep learning model and used Otsu's thresholding method for the segmentation of sperm head and other methods for the enhancement of images with accuracy reached 97.92% [17].

## III. MATERIAL AND METHODS

This section is divided into two parts, description of sperm head dataset, preprocessing methods of sperm head images with showing the impact of these methods on images. In the first part, all information about the dataset. In the second part, the main preprocessing methods for images will be explained.

### A. Dataset

The Dataset is free for using, used by McCallum et al [15]. The dataset was processed before using, and then classified into normal and abnormal sperm heads with the aid of andrologists. Dataset is divided into test dataset, validation dataset and training dataset. The dataset contains 1200 images which divided into 240 images for testing, 240 images for validation, and 720 images for training, the dataset divided into 2 classes as healthy and unhealthy sperm as represented in Fig. 1.

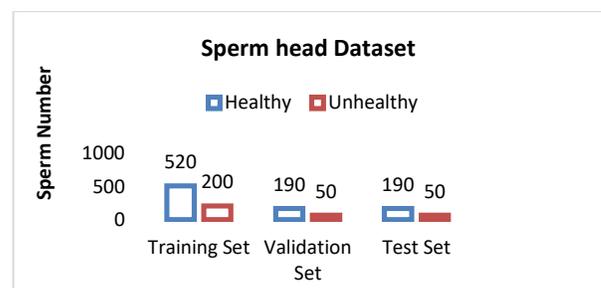


Fig. 1. Number of Normal and Abnormal Sperm Head Dataset.

## B. Preprocessing

Image preprocessing through this research is explained as the following:

1) *Image denoising*: In this step, the median filter is used for removing the salt and pepper noise for all images for making the images more suitable for the proposed method.

2) *Image normalization*: This method is used for changing the range for values of pixel intensity in image for getting a better contrast image, this process is used for adjusting the contrast of sperm head images as shown in the image Fig. 2. The linear normalization of an image is represented as the following formula:

$$O = \frac{255*(i-n)}{(m-n)} \quad (1)$$

where O is the output channel and it is the output value which is calculated, I is the input channel, n is the lowest value for pixel intensity and m is the highest value for pixel intensity.

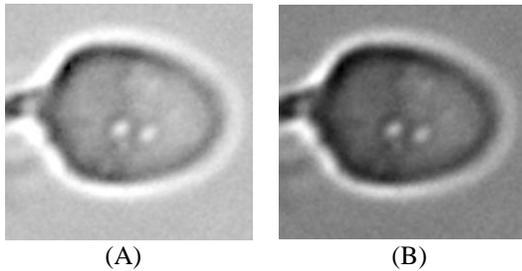


Fig. 2. Pre-processing of images. (A) Image before normalization. (B) Image after normalization.

3) *Image augmentation*: Data augmentation is a method used usually for changing the shape of images and make them little different using versatile parameters that modify the images for increasing the number of training images. Data augmentation of images is done while the training of sperm images. This is performed by ImageDataGenerator class of deep learning Keras library with using different augmentation parameters. The data augmentation is so important for increasing the accuracy of validation and training curves and for decreasing the loss of validation and training curves while training through epochs. The data augmentation contribute to minimize the overfitting.

## IV. THE PROPOSED RESNET-50 MODEL

The architecture ResNet-50 is one of the powerful deep neural networks which is used for classification tasks [16]. The ResNet-50 is used in this paper for classification of sperm heads for the best sperm detection that can be utilized in the injection into oocytes during performing the ICSI process. The convolution process is the method in which the two arrays of same or different sizes are multiplied to be merged giving third result, but it is necessary that the two arrays have the same dimensions. The mask or the matrix of convolution used may be called 'Kernel' and it is used to move over the convolution getting the required information with fewer dimensions and help in the convolution operation.

We can manage the image using this operation and get results with many processing operations for images including image sharpening, image blurring, controlling the contrast of the image and decreasing and increasing the brightness of the image.

The resulted image of this process is resulted from the rolling of the mask over all pixels of image and the final result of the multiplications are combined together and a particular calculation operation is applied resulting new pixel value that takes place in the new image and these steps are repeated for all pixels.

In this study, the better sperms can be identified utilizing ResNet50. In the beginning, the images were collected, some operation methods are important before accessing by ResNet-50. The proposed ResNet-50 model was effective in this process and it is composed of layers of convolution followed by some layers as shown in Fig. 3. The salt and pepper noise was eliminated by median filter and the training and evaluation of model should be done. Therefore, any unknown image will go through preprocessing steps after insertion for the detection. The preprocessing stages contribute to increasing the rate of success of this model then the success of the detection method with high rates. The convolution algorithm is necessary to be implemented using the convolution equation as the following:

$$O(u, v) = (F * C)(u, v) = \sum_x \sum_z M(u - X, v - Z)C(x, z) \quad (2)$$

where C is the convolution kernel of size x, O is the feature map of the output, F is the feature map of the input and (u, v) is the size of image.

Convolution steps can be described as:

- Multiplying the filter with the input image using the convolution equation.
- Multiplying each element with the element in the position then the results will be summed producing one output value.
- These steps will be reiterated by sliding the filter over the image for obtaining the output.

Average pooling in this proposed model is used for down sampling feature maps by computing the average value in each patch of a feature map. Fully connected is a layer which has multiple neurons. Dropout layer and batch normalization used for avoiding the overfitting problems. The sigmoid function is the output function and it is used for sperm head detection. The sigmoid function can be represented as the following,

$$S(x) = \frac{1}{1 + e^{-x}} \quad (3)$$

where S(x) is the sigmoid function of a variable x

The rectified linear unit (ReLU) is an activation function that makes the output like the input directly if it is positive, otherwise, makes the output zero. The ReLU activation function can be given as the following:

$$f(x) = \max(0, x) \quad (4)$$

where f(x) is the ReLU activation function of a variable x.

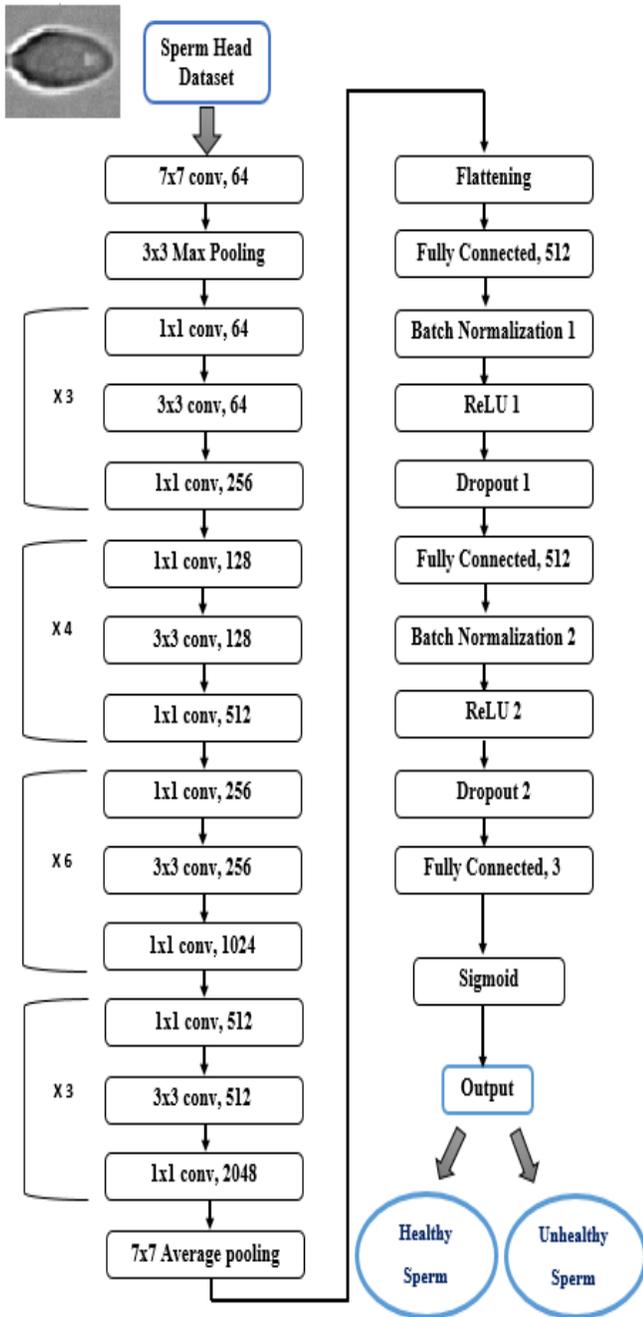


Fig. 3. The proposed resnet-50 model for healthy sperm detection.

### V. RESULTS AND DISCUSSION

The results of this proposed model demonstrated their effectiveness for the detection of normal sperms that can be utilized for the ICSI process with high accuracy. The proposed ResNet-50 model is an efficient deep learning model for classification of images and proved that in this study. In this paper, we got 165 TP value, 5 FN value, 3 FP value and 67 TN value as shown in the confusion matrix in Table I. The proposed model proved its efficiency with high accuracy reaches 96.66%, sensitivity equals 97.06%, precision equals 98.21%, Specificity equals 95.71% and F1 Score equals

97.63% as shown in Table II according to the formulas [5]–[9]. The accuracy considered as the fraction of total results that is predicted correctly to the overall number of samples. The precision is known as the proportion of values of true positive to overall predicted positive values, the sensitivity is the proportion of true positive values to total actual positive values, and the specificity is the proportion of values of true negative to the overall negative values. The detection results of this model are Normal or Abnormal as shown in Fig. 4 for the inserted unknown images. The performance of the proposed model is shown in Fig. 5 with training and validation accuracy respectively reached 97.82% and 98.41% after training with 120 epochs. The training and validation loss respectively reached 1.4026 and 1.5254. The ResNet-50 proposed model in this study is more suitable for sperm detection than other models with simpler preprocessing stages. The advantage of this proposed model is the possibility of getting accurate results quickly for making the ICSI process much easier for the andrologists.

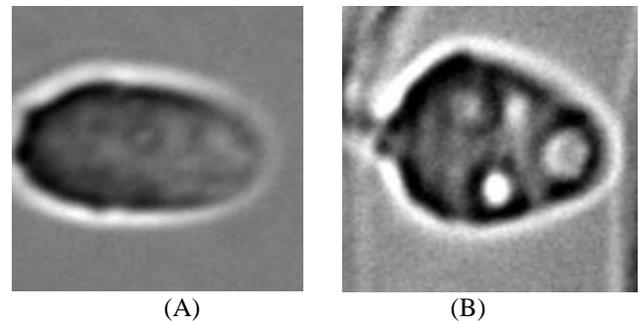


Fig. 4. Results of unknown images. (A) Normal sperm result. (B) Abnormal sperm result.

$$Accuracy = \frac{TN+TP}{TN+TP+FP+FN} \quad (5)$$

$$Sensitivity = \frac{TP}{TP+FN} \quad (6)$$

$$Precision = \frac{TP}{TP+FP} \quad (7)$$

$$Specificity = \frac{TN}{TN+FP} \quad (8)$$

$$F1\ Score = \frac{2 * Precision * Sensitivity}{(Precision+Sensitivity)} \quad (9)$$

TABLE I. THE CONFUSION MATRIX OF PROPOSED RESNET-50 MODEL

		Predicted	
		Healthy	Unhealthy
Actual	Healthy	TP = 165	FN = 5
	Unhealthy	FP = 3	TN = 67

TABLE II. EVALUATION PARAMETERS FOR THE PROPOSED RESNET-50 MODEL

Classifier	Accuracy	Sensitivity	Precision	Specificity	F1 Score
Proposed Method	96.66%	97.06%	98.21%	95.71%	97.63%

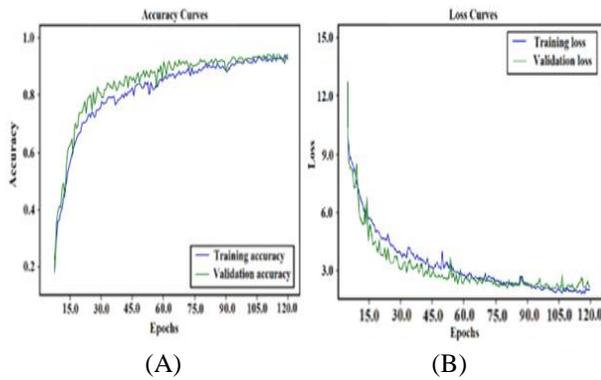


Fig. 5. Performance of the proposed resnet-50 model. (A) Training and validation accuracy curves. (B) Training and validation loss curves.

## VI. CONCLUSION AND FUTURE WORK

In this proposed model, the normal human sperms needed for the intracytoplasmic sperm injection process are recognized automatically. Andrologists are interested in easier tools for the optimum selection of healthy sperms for injection process into eggs. This study proved a technique using ResNet-50 and its outputs have been accepted and approved by the embryologists. In this paper, ResNet-50 proved the efficiency of the recognition of normal sperm and that is better than machine learning. The proposed deep learning model ResNet-50 resulting in high accuracy of 96.66% and this is one of the advantages of this study rather than the other model results in relative studies part. The ResNet-50 model has given its advantages for the detection of healthy sperms with automatic system in a quick and accurate way. The technology in methods of sperm selection is needed for the recognition of the best sperms that are required for raising success rate of ICSI operations. This system proved its distinction for obtaining good results with better accuracy for obtaining the normal sperms which referring to the normal sperm heads that can be utilized through ICSI process. One of the advantages of this study is its ability to deal with non-stained images that is similar to the captured images from microscope. Most of studies deal with stained images that are not applicable because the stain is harmful for the human sperms and make them not suitable for the injection process.

In the future, various models of deep learning can be utilized for achieving more accurate results. Developing algorithms for the detection of sperms using videos which can be built in microscope for real time detection of live sperms.

## REFERENCES

- [1] Isidori, Aldo, Maurizio Latini, and Francesco Romanelli. "Treatment of male infertility." *Contraception* 72.4 (2005): 314-318.
- [2] World Health Organization. "WHO laboratory manual for the examination and processing of human semen." (2010).
- [3] Farahani, Linda, et al. "The semen microbiome and its impact on sperm function and male fertility: a systematic review and meta-analysis." *Andrology* 9.1 (2021): 115-144.
- [4] Riordon, Jason, Christopher McCallum, and David Sinton. "Deep learning for the classification of human sperm." *Computers in biology and medicine* 111 (2019): 103342.

- [5] Javadi, Soroush, and Seyed Abolghasem Mirroshandel. "A novel deep learning method for automatic assessment of human sperm images." *Computers in biology and medicine* 109 (2019): 182-194.
- [6] Mirroshandel, Seyed Abolghasem, and Fatemeh Ghasemian. "Automated morphology detection from human sperm images." *Intracytoplasmic sperm injection*. Springer, Cham, 2018. 99-122.
- [7] Miah, Erfan, Seyed Abolghasem Mirroshandel, and Alexis Nasr. "Genetic Neural Architecture Search for automatic assessment of human sperm images." *Expert Systems with Applications* 188 (2022): 115937.
- [8] Prabakaran, L., and A. Raghunathan. "An improved convolutional neural network for abnormality detection and segmentation from human sperm images." *Journal of Ambient Intelligence and Humanized Computing* 12.3 (2021): 3341-3352.
- [9] Hidayatullah, P., et al.: "Deepsperm: a robust and real-time bull sperm-cell detection in densely populated semen videos." *Comput. Methods Programs Biomed.* 209 (2021).
- [10] Hicks, S., et al. "P-272 Automatic Tracking of the ICSI procedure using Deep Learning." *Human Reproduction* 37.Supp1\_1 (2022): deac107-261.
- [11] Marín, Ruth, and Violeta Chang. "Impact of transfer learning for human sperm segmentation using deep learning." *Computers in Biology and Medicine* 136 (2021): 104687.
- [12] Lee, Ryan, et al. "Automated rare sperm identification from low-magnification microscopy images of dissociated microsurgical testicular sperm extraction samples using deep learning." *Fertility and Sterility* (2022).
- [13] Golomingi, Raffael, et al. "Sperm hunting on optical microscope slides for forensic analysis with deep convolutional networks—a feasibility study." *Forensic Science International: Genetics* 56 (2022): 102602.
- [14] Aristoteles, Aristoteles, et al. "Identification of Human Sperm based on Morphology Using the You Only Look Once Version 4 Algorithm." *International Journal of Advanced Computer Science and Applications* 13.7 (2022): 424-431.
- [15] McCallum, Christopher, et al. "Deep learning-based selection of human sperm with high DNA integrity." *Communications biology* 2.1 (2019): 1-10.
- [16] Koonce, Brett. "ResNet 50." *Convolutional neural networks with swift for tensorflow*. Apress, Berkeley, CA, (2021): 63-72.
- [17] Mashaal, A.A., Eldosoky, M. A., Mahdy, L. N., & Kadry, A. E. "Automatic Healthy Sperm Head Detection using Deep Learning." *International Journal of Advanced Computer Science and Applications*, 13.4 (2022): 735-742.
- [18] Baker, Bowen, et al. "Designing neural network architectures using reinforcement learning." *arXiv preprint arXiv:1611.02167* (2016).
- [19] Zoph, B., Le, Q.V. "Neural architecture search with reinforcement learning." *arXiv preprint arXiv:1611.01578* (2016).
- [20] Schubert, Benoit, Mélanie Badiou, and André Force. "Computer-aided sperm analysis, the new key player in routine sperm assessment." *Andrologia* 51.10 (2019): e13417.
- [21] Guzick, David S., et al. "Sperm morphology, motility, and concentration in fertile and infertile men." *New England Journal of Medicine* 345.19 (2001): 1388-1393.
- [22] MORTIMER, DAVID, and ROELOF MENKVELD. "Sperm morphology assessment—historical perspectives and current opinions." *Journal of andrology* 22.2 (2001): 192-205.
- [23] Kumar, Naina, and Amit Kant Singh. "Impact of environmental factors on human semen quality and male fertility: a narrative review." *Environmental Sciences Europe* 34.1 (2022): 1-13.
- [24] Baskaran, Saradha, et al. "Diagnostic value of routine semen analysis in clinical andrology." *Andrologia* 53.2 (2021): e13614.
- [25] Esteves, Sandro C. "Evolution of the World Health Organization semen analysis manual: where are we?." *Nature Reviews Urology* (2022): 1-8.

# Research on Image Sharpness Enhancement Technology based on Depth Learning

Wenbao Lan\*, Chang Che

School of Civil Engineering, Harbin University, Harbin, 150086, China

**Abstract**—Image technology is widely used in security, traffic, monitoring, and other social activities. However, these images carrying detailed information will have feature distortion due to various external physical factors in social ingestion, transmission, and storage, resulting in poor image quality and clarity. Resolution determines the definition of an image. Super-resolution reconstruction is the process of transforming a low-resolution picture into a high-resolution image. To enhance the image clarity, this experiment introduces the advantages and disadvantages of the Super Resolution Convolutional Network (SRCNN) and Fast Super Resolution Convolutional Neural Network (FSRCNN) model and then constructs an image super resolution method based on DSRCNN. The algorithm consists of two sub-network blocks, an enhancement block and a purification block. The model first uses two Convolutional Neural Networks (CNN) to obtain complementary low-frequency information that improves the model's learning ability; next, it employs an enhancement block to fuse the image features of two paths via residual operation and sub-pixel convolution to prevent the loss of low-resolution image information; finally, it employs a feature purification block to refine high-frequency information that more accurately represents the predicted high-quality image. It is found that the PSNR and SSIM of the DSRCNN model can reach 33.43dB and 0.9157dB, respectively.

**Keywords**—Super resolution convolutional network; fast super resolution convolutional neural network; dual super-resolution convolutional network; deep learning; image definition; super resolution; image enhancement

## I. INTRODUCTION

Pictures captured in social life are often affected by the light, shelter, poor shooting equipment, and other factors. Such shootings have poor visibility, low contrast, high noise, and other shortcomings [1]. These shortcomings make the image definition fail to reach the ideal state, which affects the regular operation of some links to social activities. Picture sharpness is directly related to the image resolution. High-resolution images have more detail and are of superior quality [2]. Processing a series of low-resolution images to obtain high-resolution images is called super-resolution reconstruction. Traditional image super-resolution reconstruction is mainly based on interpolation, machine learning, and multiple images. Researchers value the super-resolution reconstruction technology based on deep learning algorithms because of the ongoing changes in computing technology [3]. The deep learning algorithm mainly learns the high-level semantic features of data through the transformation of multi-level nonlinear structure. It can achieve reasonable and effective prediction and analysis of data [4]. Classical depth learning algorithms include SRCNN, FSRCNN, VDSR, etc. These

super-resolution reconstruction technologies can improve image clarity, but there are still some limitations in operation efficiency and visual effects [5]. Therefore, this experiment suggests an image definition improvement technique based on deep learning algorithms to improve the picture quality with inadequate definition, aiming to add reference value to image applications in social activities such as security, traffic, monitoring, etc.

## II. RELATED WORK

At present, the application of depth learning is very broad, and it has been dabbled in various research fields. Wang S et al. used depth learning algorithm to recognize tumor images, and introduced the whole process of depth learning algorithm in pathological image segmentation in detail. The research shows that the deep learning algorithm has obvious advantages in recognition accuracy, overall computational efficiency and generalization performance. In addition, integrating in-depth learning into the field of pathological image analysis is a far-reaching experiment in medicine [6]. Huang and other researchers found that traditional machine learning methods had limited prediction ability for landslide prone environmental factors, so they proposed a FC-SAE algorithm based on deep learning. To predict categorization, the system employs both high-level and low-level information about environmental elements. The simulation results show that the prediction rate and accuracy rate of FC-SAE can reach 85.4% and 85.2%, higher than other algorithms of the same type, so it has a good application prospect [7]. Sharma and other scholars explored deep neural networks based on machine learning. For the goal of conducting research on the use of machine learning and deep learning in different sectors, they examined and improved the current machine learning and deep learning algorithms. The final survey results show that the accuracy of deep learning algorithms has been significantly improved [8]. Zhang and other researchers conducted research on deep learning algorithms in the field of geological engineering. This paper mainly studies the application of FNN, RNN, CNN and GAN depth learning algorithms in geotechnical engineering, compiles a detailed summary containing literature, cases and depth learning algorithms, and prospects the application of depth learning algorithms in geotechnical engineering [9]. Aslan and others found that malware is constantly developing and seriously endangering the network security environment. Based on this malignant situation, they propose to use deep learning algorithm to detect malicious software. The suggested deep learning algorithm is a novel hybrid architecture that combines two comprehensive pre-training network models, which makes it different from conventional approaches. The

experimental findings demonstrate that the model can classify malware with an accuracy of 97.78% [10]. Bamisile and other scholars used machine learning algorithm and depth learning algorithm to predict solar radiation. SVR, RF, and their learning models as well as ANN, CNN, and RNN deep learning models are each utilized in the experiment to forecast solar radiation, and the outcomes of these predictions are compared. The final results show that the deep learning model has better accuracy and takes less time than the machine learning model in solar radiation prediction [11].

For the research of image enhancement technology, most researchers have made contributions to it. An underwater image enhancement approach based on MLE was suggested by Zhang and other researchers to address the issues of poor color and low visibility of underwater photos. The method first adjusts the color and detail of the image by the principle of minimum color loss, then uses the integral graph to calculate the mean and variance of the image block, and finally introduces a color balance strategy to balance the color difference of the image. The outcomes of the experiment demonstrate that this technique is very applicable to picture enhancement [12]. Ike C and other researchers proposed a single image super-resolution method based on wavelet. The experiment combines wavelet transform with local regularization anchored domain regression model and applies it to image restoration. Experimental results show that the proposed method is effective and superior in denoising, deblurring and super-resolution reconstruction tasks [13]. Zhang and other scholars proposed a method that can enhance the image contrast while improving the embedded data, namely the improved RDH-CE model. The model offers an adaptive pixel modification technique and extracts several characteristics to aid K-means clustering. According to the final testing findings, this technique has a greater accuracy level for complicated assessment [14]. Liu and other researchers used MSRCR algorithm and guided filtering method to de fog the image. To provide a better visual impact, this approach may successfully address the issues of poor picture lighting, color distortion, and edge loss while preserving the color saturation of the image [15]. A band restricted biphasic technique was suggested by Su et al. to enhance the quality of the reconstructed picture. By using this technique, the precise spatial frequency components impacted

by the spatial offset of DPHs may be removed, preventing the picture quality from being compromised. According to the outcomes of the simulation experiment, the strategy enhances the definition of the rebuilt picture by 36.84% [16]. To solve the serious distortion problem of underwater images, Liu et al. proposed an image enhancement method based on object guided dual confrontation contrast learning, which uses contrast prompts in the training phase. The experiment demonstrates that the suggested strategy enhances the image's visual quality and increases the detector's precision [17].

Deep learning algorithm is favored by researchers in various fields for its convenient and powerful performance, and its application is also more extensive. In recent years, the research on image enhancement technology has never stopped, and many scholars have used different algorithms to explore it. In conclusion, deep learning algorithm and image enhancement technology have made considerable progress. It is an innovative attempt to apply the depth learning algorithm to the image definition enhancement technology, so this experiment starts to study this topic.

### III. CONSTRUCTION OF IMAGE DEFINITION ENHANCEMENT MODEL BASED ON DEPTH LEARNING ALGORITHM

#### A. Research on Image Super-resolution Model based on Depth Learning

The resolution of an image determines its definition, and super resolution is the resolution of an image that has been enhanced using a variety of efficient techniques. Super Resolution refers to the technique of processing a number of low-resolution photographs to produce high resolution images (SP), which mainly includes the reconstruction of a single image and the reconstruction of multiple images [18-19]. Super Resolution Convolutional Network (SRCNN) is a milestone for depth learning algorithm to achieve super-resolution image reconstruction [20]. SRCNN uses a three-layer convolutional neural network to achieve image reconstruction, that is, to preprocess the low-resolution image through cubic interpolation. Fig. 1 illustrates the three components that make up the SRCNN model, namely block precipitation and representation, nonlinear mapping and reconstruction. The three modules correspond to three convolution operations in the network.

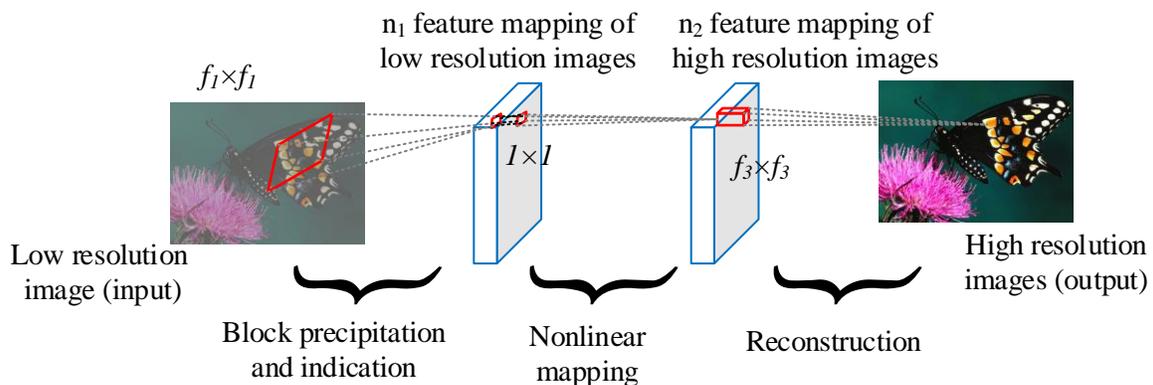


Fig. 1. Network structure of SRCNN.

$B$  represents low resolution image;  $A$  represent real high-resolution images; the mapping function of the network is  $f(\cdot)$ . Consequently, the equation for block precipitation and module expression in SRCNN is as follows:

$$f_1(B) = \max(0, W_1 * B + \delta_1) \quad (1)$$

In Formula (1),  $W_1$  represents the weight of the convolution core of the block precipitation module;  $\delta_1$  indicates the offset of the convolution core of the block precipitation module.  $\max$  represents the corresponding ReLU activation function;  $f_1(B)$  represents block precipitation. The nonlinear mapping of the network is called "convolution+activation", and its functional expression is:

$$f_2(B) = \max(0, W_2 * f_1(B) + \delta_2) \quad (2)$$

In Formula (2),  $W_2$  represents the weight of the nonlinear mapping module's convolution kernel,  $B_2$  represents the offset of the convolution kernel of the nonlinear mapping module;  $f_2(B)$  represents a nonlinear mapping. The network reconstruction process is the third convolution operation. Unlike the two modules mentioned above, the activation function is no longer used in this process. Its function expression is:

$$f(B) = W_3 * f_2(B) + \delta_3 \quad (3)$$

In Formula (3),  $W_3$  represents the weight of the convolution core of the reconstruction module,  $\delta_3$  represents the offset of the convolution core of the reconstruction module, and  $f(B)$  represents the reconstruction. The advantage of SRCNN model is that it contains fewer parameters than other deep convolution neural networks. As a result, by using fewer learning parameters, it may significantly minimize the number of calculations required by the model, thus improving the overall model operation efficiency. The loss function of SRCNN is the mean variance between the super resolution image  $f(B)$  and the original real high resolution image  $A$ . The specific calculation formula is as follows:

$$L(\theta) = \frac{1}{n} \sum_{i=1}^n \|f(B_i; \theta) - A_i\|^2 \quad (4)$$

In Formula (4),  $n$  represents the number of samples in the training set,  $f(B_i, \theta)$  represents the minimum reconstructed image, and  $A_i$  represents the high-resolution image. The loss of mean variance represents the loss mean of a batch, which is an evaluation index of commonly used image high-resolution algorithms. Although the classic SRCNN network has been able to repair low resolution images with different coefficients, the SRCNN network still has limitations that cannot meet the real-time requirements. Thus, FSRCNN network came into being. FSRCNN network consists of five parts: feature extraction, compression, mapping, expansion and transpose

convolution [21-22]. The FSRCNN network is used to extract and compress the image features, and then the activation function pair is used for nonlinear mapping. The activation function of FSRCNN may activate the gradient of the negative half axis in addition to the SRCNN's activation function. The precise formula may be written as follows:

$$g(x_i) = \max(x_i, 0) + a_i * \min(0, x_i) \quad (5)$$

In Formula (5),  $x_i$  represents the input signal of the activation function in the  $i$ -th channel;  $a_i$  represents the coefficient of the negative part. The activation function has more stable performance. The calculation complexity of FSRCNN model can be expressed as:

$$O\left\{\left(25d + sd + 9ms^2 + ds + 81d\right)S_{LR}\right\} = O\left\{\left(9ms^2 + 2sd + 106d\right)S_{LR}\right\} \quad (6)$$

In Formula (6),  $d$  represents the characteristic layer before compression;  $s$  represents the compressed feature layer; And both meet  $s < d$ .  $m$  indicates the number of layers of the volume. It can be seen from Formula (6) that the computational complexity of FSRCNN model is in linear proportion to LR image, and its computational complexity is also less than SRCNN. The loss function calculation formula of FSRCNN can be expressed as:

$$\min_{\theta} \frac{1}{n} \sum_{i=1}^n \|f(B_s^i; \theta) - A^i\|_2^2 \quad (7)$$

In Formula (7),  $A^i$  represents the  $i$ -th high-definition image in the data set;  $B_s^i$  represents the  $i$ -th resolution image in the dataset.  $f(B_s^i; \theta)$  represents the output of the model. Compared with SRCNN network, FSRCNN network is mainly improved in three aspects, namely, the change of feature dimension, the replacement of convolution core with more mapping layers and the sharing of mapping layers. Although SRCNN and FSRCNN both improve the quality of super resolution images, they are generally slow and not suitable for real application scenarios. The research on the application of depth learning algorithm to the improvement of super-resolution image quality needs to be further deepened.

### B. Improvement of Image Definition Enhancement Model based on Fusion of Depth Learning Algorithm

In this experiment, a dual super resolution CNN (DSRCNN) is designed based on SRCNN and FSRCNN models. The network consists of two sub network modules, enhancement blocks and feature purification blocks. Two sub network blocks enhance the super resolution performance of the network by extracting complementary low-frequency features; the enhancement block gathers several high-frequency characteristics via residual operation and sub-pixel convolution; the feature purification block uses multiple stacked convolutions to refine high-frequency features [23-24]. The process of extracting complementary low-frequency features from two sub network blocks of DSRCNN model is as follows:

$$O_{TSEB} = f_{TSEB}(I_{LR}) = \left( Cat(O_{TSEB_1}, O_{TSEB_2}) \right) \quad (8)$$

In Formula (8),  $O_{TSEB_1}$  and  $O_{TSEB_2}$  represent the output of two subnetworks, and  $Cat$  represents the connection operation. In the second stage, the two subnetworks select the combination of reusing residual operation and convolution to enhance the effect of local level information. The formula can be expressed as:

$$O_{TSEB_k} = R(O_{L1} + O_{L2} + O_{L5} + O_{L7} + O_{Li} + \dots) \quad (9)$$

In Formula (9),  $O_{(TSEB_k)}$  represents the output of  $k$  sub network,  $O_{Li}$  represents the output of the  $i$ -th convolution layer, and  $R$  represents ReLU function. The purpose of the model's enhancement block is to combine the original picture data and gain deeper data. This module first obtains the original image and deeper information, and obtains the characteristics of the first layer in the sub network:

$$\begin{aligned} O_{(EB_1)} &= Subp(R(O_{L1})) \\ O_{(EB_i)} &= Subp(O_{TSEB}) \end{aligned} \quad (10)$$

In Formula (10),  $O_{(EB_1)}$  stands for the shallow layer's high-frequency properties, and  $O_{(EB_i)}$  for the deep layer's high-frequency features.  $Subp$  represents sub-pixel convolution technology. Then, the second level features are obtained from the features in the residual operation fusion formula (10), and its functional expression is:

$$O_{EB} = f_{EB}(O_{TSEB}) = R(O_{(EB_1)} + O_{(EB_i)}) \quad (11)$$

In Formula (11),  $O_{(EB)}$  represents the output of the enhancement block, serves as the purification block of the feature,  $+$  represents a residual operation, and  $f_{EB}$  represents the function of the enhancement block. The feature purification block equation can be expressed as:

$$I_{SR} = f_{FLB}(O_{EB}) = C\left(R\left(C\left(R\left(C\left(R\left(C\left(R\left(C\left(O_{EB}\right)\right)\right)\right)\right)\right)\right)\right)\right) \quad (12)$$

Formula (12) constructs a five layer feature purification block, which contains four convolution operations and activation functions, and can be used to build high-quality predicted images. In Formula (12),  $C$  represents the convolution operation, and  $I_{SR}$  represents the predicted SR image. The DSRCNN's execution procedure may be stated as follows:

$$I_{SR} = f\left(f_{EB}\left(f_{TSEB}\left(I_{LR}\right)\right)\right) = f_{DSRCNN}\left(I_{LR}\right) \quad (13)$$

In Formula (13),  $f_{TSEB}$  represents the function of two subnets,  $f_{EB}$  represents the function of enhancement block,  $f_{FLB}$  represents the function of feature purification block, and

$I_{LR}$  represents the given LR image. DSRCNN mainly obtains complementary low-frequency information through dual CNN to improve the learning ability of the model. Then, to prevent the loss of low resolution picture information, the enhancement block merges the image characteristics of two separate image routes using residual operation and sub-pixel convolution. In order to properly reflect the anticipated high-quality picture, high-frequency information is refined using the feature purification block [25]. The above process is described in detail in Fig. 2.

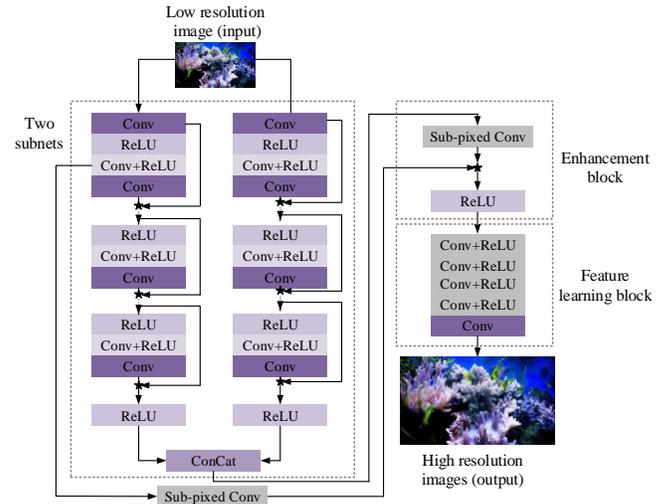


Fig. 2. Network architecture of DSRCNN.

The difference between the actual high-quality picture and the anticipated SR image serves as the mean square error, which is how the loss function of DSRCNN is still stated. The specific function expression is:

$$l(p) = \frac{1}{2N} \sum_{(k=1)}^N \left\| f_{DSRCNN}(I_{LR}^k - I_{HR}^k) \right\|^2 \quad (14)$$

In Formula (14),  $I_{LR}^k$  represents a given low resolution image,  $I_{HR}^k$  represents a given high resolution image,  $N$  represents the total number of training samples, and  $p$  represents the parameter set for training a DSRCNN. The value of  $l(p)$  is calculated from the training pair  $\{I_{LR}^k, I_{HR}^k\}$ . PSNR and SSIM can be selected as quantitative and objective evaluation indicators for model evaluation. Suppose the length and width of a noise image are  $M$  and  $N$  respectively,  $K$  represents the noise image,  $I$  represents the clean image, and their mean variance formula is:

$$MSE = \frac{1}{mn} \sum_{i=0}^{m-1} \sum_{j=0}^{n-1} [I(i, j) - K(i, j)]^2 \quad (15)$$

In Formula (15),  $i$  represents the feature in the  $i$ -th column of the picture,  $j$  represents the feature in the  $j$ -th column of the picture, and the formula of PSNR can be expressed as:

$$PSNR = 10 \cdot \log_{10} \left( \frac{MAX_I^2}{MSE} \right) \quad (16)$$

PSNR is measured in dBs. The less the value is, the less the picture will be warped. The PSNR index, which bases its judgment of picture quality on the associated pixel error, ignores the visual properties of the human eye. Because the human eye has a high visual sensitivity to space, brightness, color, etc., the evaluation results of PSNR are often inconsistent with the subjective feelings of the human eye. Structural similarity requires difference processing for target images  $X$  and  $Y$ .  $\mu_x$  and  $\mu_y$  respectively stand for the mean values of images  $X$  and  $Y$ ,  $\sigma_x$  and  $\sigma_y$  respectively stand for the variances of images  $X$  and  $Y$ , and  $\sigma_{xy}$  stands for the covariance of images  $X$  and  $Y$ . The formula is as follows:

$$\begin{aligned} \mu_x &= \frac{1}{M \times N} \sum_{i=1}^M \sum_{j=1}^N X(i, j) \\ \sigma_x &= \left( \frac{1}{M \times N - 1} \sum_{i=1}^M \sum_{j=1}^N (X(i, j) - \mu_x)^2 \right)^{\frac{1}{2}} \\ \sigma_{xy} &= \frac{1}{M \times N - 1} \sum_{i=1}^M \sum_{j=1}^N (X(i, j) - \mu_x)(Y(i, j) - \mu_y) \end{aligned} \quad (17)$$

SSIM assesses how similar pictures are in terms of structure, contrast, and brightness. The contrast formula is as follows:

$$\begin{aligned} l(X, Y) &= \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1} \\ c(X, Y) &= \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2} \\ s(X, Y) &= \frac{\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3} \end{aligned} \quad (18)$$

In Formula (18),  $c_1$ ,  $c_2$  and  $c_3$  are three constants, and there is no coefficient of any variable between them, but they cannot be removed. The formula of SSIM can be expressed as:

$$SSIM(x, y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \quad (19)$$

SSIM has a value range of 0 to 1. The picture distortion decreases with increasing value, and image quality and clarity increase. In the calculation process, it is also necessary to cut the global pixels of the image, and the structural similarity obtained is the global structural similarity.

#### IV. PERFORMANCE VERIFICATION OF IMAGE DEFINITION ENHANCEMENT MODEL BASED ON DSRCNN

To more accurately verify the image super-resolution reconstruction performance of the model, four data sets, Set5, Set14, BSD100 and Urban100, were selected for the experiment, and the contrast experiments were carried out under the conditions that the image magnification coefficient

was 2 times, 3 times and 4 times. Set5 consists of five images: children, birds, butterflies, flowers and women. 14 photos make up the data set known as ‘‘Set14,’’ which is often used to evaluate the effectiveness of image super-resolution models. The BSD100 dataset contains 100 daily life events, which are often used for image denoising and super-resolution testing. Urban100 is an image data set with the theme of street view, in which the images show complex color characteristics, and their objects have clear shapes and corners. The conditions under which Set5 and Set14’s photos were taken are the same. BSD100 and Urban100 both provide 100 photos in natural color. The experimental training set and test set each include 100 natural photos, and the image data in the data set DIV2K is chosen. Table I shows the settings of other experimental parameters.

TABLE I. EXPERIMENTAL PARAMETER SETTING TABLE

Experimental parameters	Set value	Experimental parameters	Set value
Batch size	64.000	Training step	600000
Parameters Beta1	0.900	Epsilon	100000000
Parameters Beta2	0.999	Learning rate of model training 1	0.0004
Initial learning rate	0.001	Learning rate of model training 2	0.00005

Fig. 3 shows the visual test results of SRCNN model, FSRCNN model and DSRCNN model on visual graphics. To create these visual graphics, select a specific area of the anticipated super-resolution image as the observation area. The data in the figure shows that the PSNR and SSIM values of SRCNN model are 27.002 and 0.928; PSNR and SSIM values of FSRCNN model are 30.783 and 0.8666; The PSNR and SSIM values of the DSRCNN model are 33.783 and 0.9688. It can be observed from the above data that in the three models, since DSRCNN’s PSNR and SSIM values are the greatest, its images exhibit the least amount of distortion. This demonstrates that the DSRCNN model performs superior to the other two models.

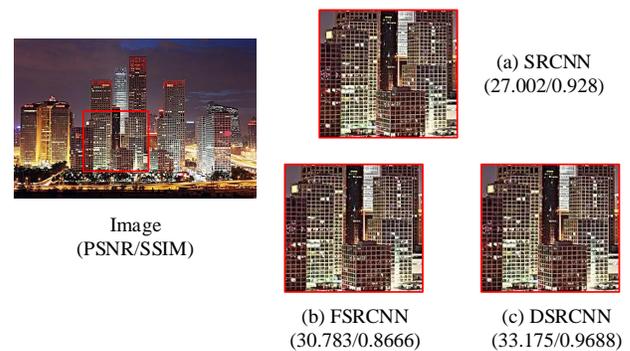


Fig. 3. Visual diagram of three models.

Table II shows the comparison of PSNR and SSIM values of different modules of DSRCNN in Urban100 dataset with twice the image magnification factor. The results in the table demonstrate that local residual learning is efficient for the development of model performance as the PSNR and SSIM values of the DSRCNN network are better than those of the DSRCNN model with residual learning method. It is observed

that the DSRCNN model is superior to the DSRCNN with only one subnet, and the two subnets in the table model contribute to the performance improvement of the model. The PSNR value of DSRCNN with enhancement block is 0.454dB higher than that of DSRCNN without enhancement block, which verifies the improvement of network performance by enhancement block. The PSNR and SSIM values of the DSRCNN model with feature purification block are 0.16dB and 0.011dB higher than those of the DSRCNN model without feature purification block, respectively, which proves that the feature purification block can improve the network performance. As a result, the three modules in the DSRCNN model proposed in the experiment all have different contributions to their own performance improvement, and this model can improve the image quality.

To better compare the performance of the DSRCNN model proposed in this paper, in addition to the comparison with the SRCNN model, the experiment also selects six other popular depth learning models, a total of 10 models, to conduct experiments to compare and analyze their performance. The popular models selected are Bicubic model, CNF model, MemNet model, CARN-M model, WaveRestNet model and LESRCNN-S model. Fig. 4 displays the variations in PSNR and SSIM values for the chosen model on the Set5 dataset at 2, 3, and 4 times the original picture size. When the image is magnified twice, the DSRCNN model can have the next highest PSNR value, 37.74dB. When the image magnification is 3 times, the DSRCNN model has the highest PSNR value, 34.17dB. When the image magnification is 4 times, the DSRCNN model achieves the highest SSIM value of 0.8907. On the whole. The DSRCNN model shows relatively stable performance on the Set5 dataset.

Fig. 5 displays the changes in PSNR and SSIM values for several models using the Set14 dataset with picture magnifications of 2, 3, and 4. When the image is magnified twice, the PSNR value of the DSRCNN model is 33.43, which is the highest among all models. At the same time, its SSIM

value is 0.9157, which is the highest among all models. When the magnification of the image is 3 times, the PSNR of the DSRCNN model is still the highest among all models, 30.24, and its SSIM value is 0.8402, which also keeps the highest record. When the image magnification is 4 times, the PSNR value of the DSRCNN model is 28.46, and the SSIM value is 0.7796. Although it has decreased, it still maintains the highest value of all models. On the Set14 dataset, it is clear that the DSRCNN model performs well.

Fig. 6 displays the variations in PSNR and SSIM values for several models on the BSD100 dataset at 2, 3, and 4 times the original picture size. When the image magnification is 2 times, the PSNR value of the DSRCNN model is 32.05 and the SSIM value is 0.8978. When the image magnification is 3 times, the PSNR value of the DSRCNN model is 29.01 and the SSIM value is 0.8029, ranking first among all models. When the image magnification is 4 times, the PSNR value of the DSRCNN model is 27.50, and the SSIM value is 0.7341, which is also in the first place. This shows that on the BSD100 dataset, the DSRCNN model still has better performance than other models.

TABLE II. COMPARISON OF PSNR VALUE AND SSIM VALUE OF DIFFERENT MODULES IN DSRCNN ON URBAN100 DATASET WITH TWICE IMAGE MAGNIFICATION FACTOR

Methods	Urban100
	PSNR/SSIM
DSRCNN with one sub-network ,without RLO	31.675/0.9236
DSRCNN with one sub-network and EB_S, without RLO,	31.221/0.9181
DSRCNN with one sub-network EB_S andRO,without RLO	31.061/0.9170
DSRCNN with EB_S , without RLO	31.648/0.9237
DSRCNN without RLO	31.700/0.9237
DSRCNN	31.834/0.9253

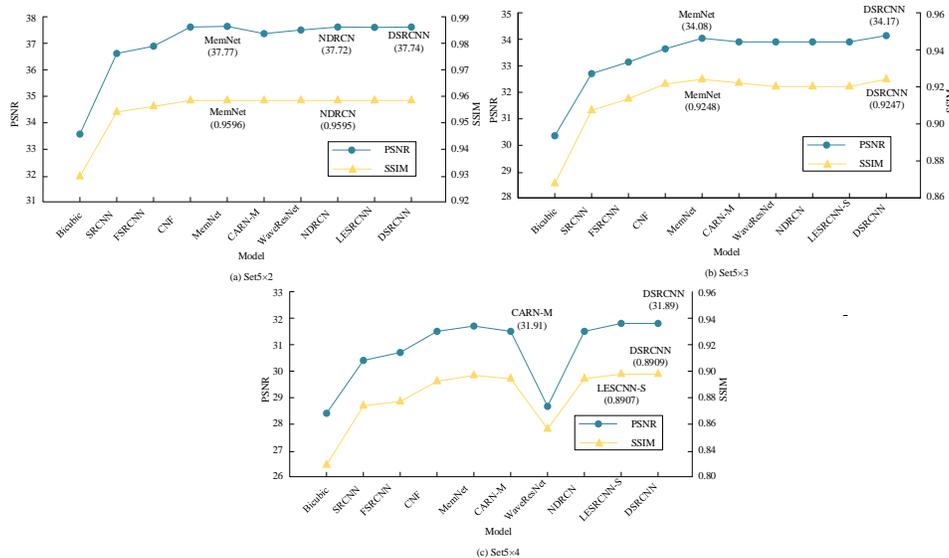


Fig. 4. Changes of PSNR and SSIM values of different models on Set5.

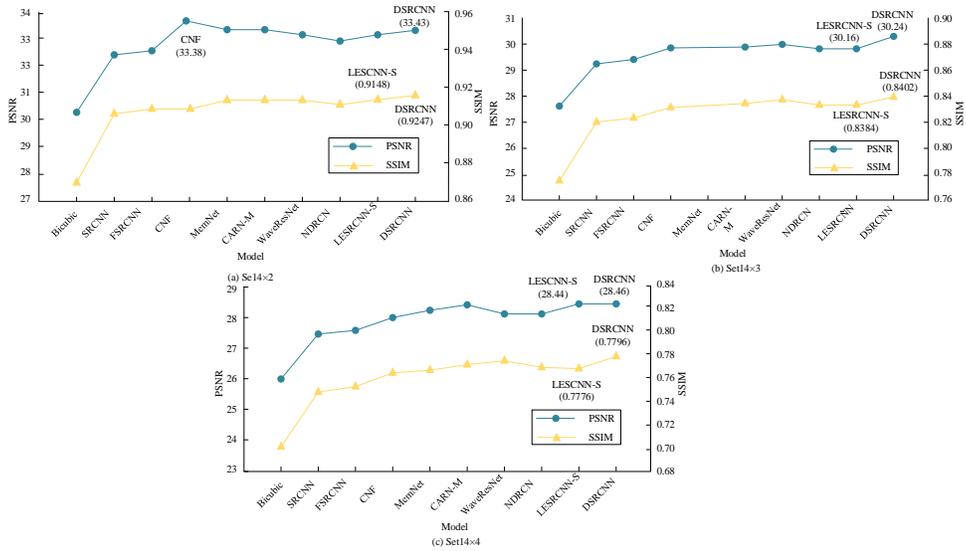


Fig. 5. Changes of PSNR and SSIM values of different models on Set14.

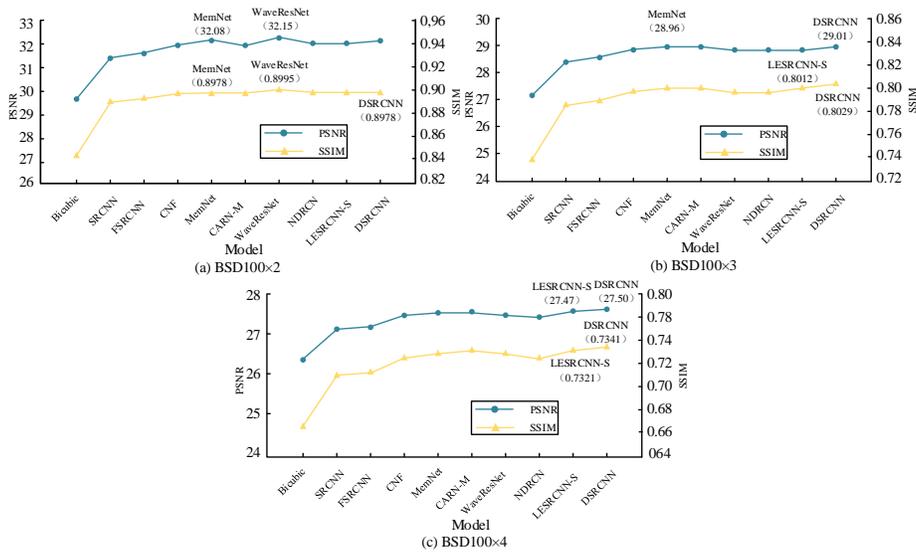


Fig. 6. Changes of PSNR and SSIM values of different models on BSD100.

Fig. 7 displays the variations in PSNR and SSIM values between models using the Urban100 dataset with picture magnifications of 2, 3, and 4. When the image magnification is 2 times, the PSNR value of the DSRCNN model is 31.83, and the SSIM value is 0.9252, which is the highest among all models. When the image magnification is 3 times, the PSNR value of the DSRCNN model is 27.76, and the SSIM value is 0.8483, ranking the highest among all models. When the image magnification is 4 times, the PSNR value of the DSRCNN model is 25.94, and the SSIM value is 0.7815, which is always the first of all models. This shows that the DSRCNN model also has good performance on the Urban100 dataset.

Fig. 8 shows that on the dataset Urban100, when the magnification factor is 4 times the visual effect of different models, the data in the figure shows that the PSNR and SSIM values are highest for the DSRCNN model. To test the effectiveness of the DSRCNN model based on the deep learning algorithm, four distinct datasets are chosen. The

results showed that the PSNR and SSIM values of DSRCNN model fluctuated in different data sets for different models, but among the 10 chosen models, image processing performance as a whole was best. This shows that the DSRCNN model can handle the image processing, making the final image less distorted, improving its separation rate, and thus enhancing the image clarity.

Table III shows the network complexity analysis results of DSRCNN and the four models selected in the experiment. It can be seen from the table that the number of parameters of DSRCNN model is much smaller than that of the other models on the Urban100 dataset with 2, 3 and 4 times of image magnification. In addition, its number of operations (Flops) is also less than other models. The comparison results further reflect the superior data processing ability and performance of DSRCNN. At the same time, it also verifies the innovation of DSRCNN for image data processing methods.

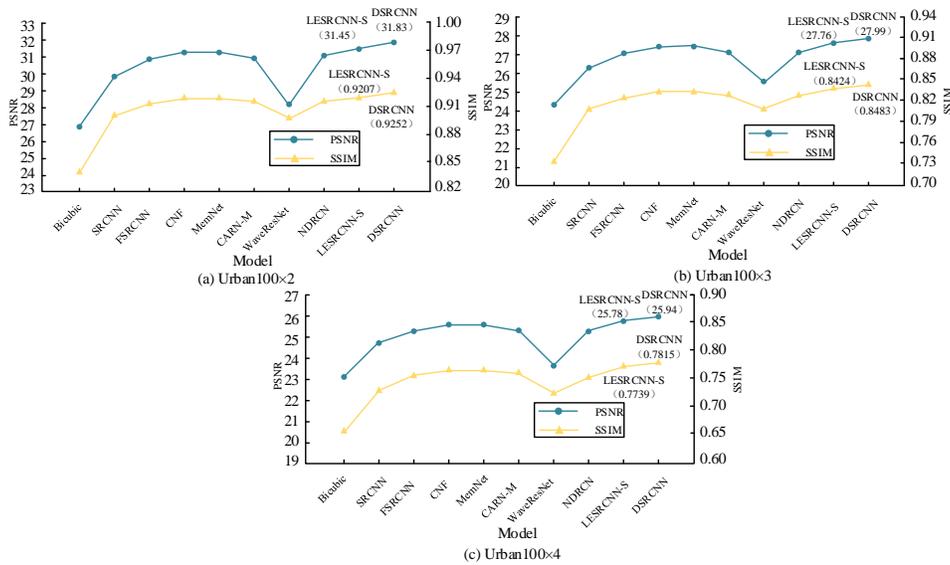


Fig. 7. Changes of PSNR and SSIM values of different models on Urban100.

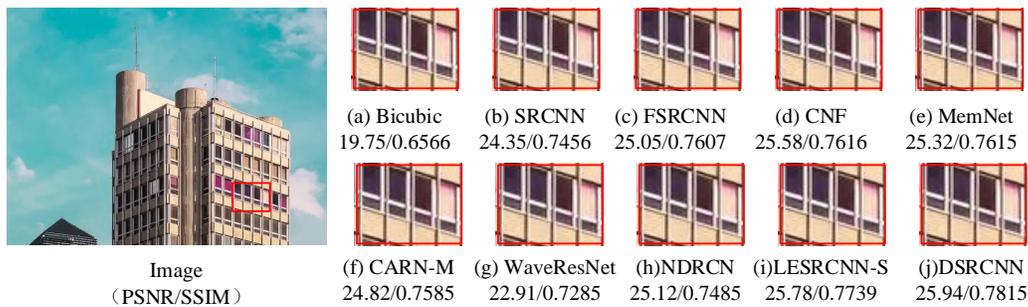


Fig. 8. Visual effects of different models.

TABLE III. COMPARISON OF MODEL PARAMETERS

Methods	Scale×2		Scale×4		Scale×6	
	Parameter quantity(K)	Flops(G)	Parameter quantity(K)	Flops(G)	Parameter quantity(K)	Flops(G)
WaveRe Net	664	10.90	675	11.23	789	12.35
LESRCNN-s	557	9.12	603	10.12	667	11.24
SRCNN	879	29.06	978	30.14	1023	31.23
FSRCNN	789	11.09	856	12.03	964	12.99
DSRCNN	550	5.56	563	6.23	574	7.02

## V. CONCLUSION

Image technology is widely used in security, traffic, monitoring, and other social activities because it can carry detailed information. To improve the distortion in the process of image acquisition, transmission and storage, the DSRCNN model for image super-resolution reconstruction is established from the perspective of image super-resolution reconstruction and depth learning technology. This experiment aims to improve the image definition. In social ingestion, transmission, and storage, image features will be distorted due to various external factors, resulting in poor image quality and clarity. In this experiment, a DSRCNN model is built for image super-resolution reconstruction from the viewpoint of image super-resolution reconstruction and depth learning techniques to

improve the definition of the picture. At first, the performance of the DSRCNN model, SRCNN model, and FSRCNN is compared. The PSNR and SSIM values of the SRCNN model are 27.002 and 0.928. The FSRCNN model is 30.783 and 0.8666, the DSRCNN model is 33.783 and 0.9688, and so on, which verifies that the DSRCNN model has superior performance. To further explore the DSRCNN model, four different data sets are selected to verify the model performance. Under various picture magnifications, the PSNR and SSIM values of DSRCNN models in various data sets likewise change. In the data set Set14, when the magnification is 2, the DSRCNN model has the highest PSNR and SSIM values, 33.43 and 0.9157, respectively. Therefore, the DSRCNN model can effectively process the image, reduce image distortion, and improve its resolution, enhancing image clarity.

This experiment basically achieved the purpose of the experiment. In the future, we can consider putting forward effective solutions to the visual shape differentiation in the DSRCNN model.

#### ACKNOWLEDGMENT

The research is supported by the key topic of the “14th five year plan” of Education Science in Heilongjiang Province in 2022: A study on the path of building a smart classroom teaching mode “golden class” with reference to German FH education experience under the background of professional certification (No. GJB1422322).

#### REFERENCES

- [1] Y. Chen, W. Chen, S. Chandra Pal, A. Saha, I. Chowdhuri, B. Adeil, S. Janizadeh, A. Dineva, X. Wang, A. Mosavi, “Evaluation efficiency of hybrid deep learning algorithms with neural network decision tree and boosting methods for predicting groundwater potential,” *Geocarto International*, vol. 37, no. 19, pp. 5564-5584, 2022.
- [2] T. S. Kumar, T. Senthil, “Construction of hybrid deep learning model for predicting children behavior based on their emotional reaction,” *Journal of Information Technology*, vol. 3, no. 1, pp. 29-43, 2021.
- [3] D. A. Neu, J. Lahann J, P. Fettke, “A systematic literature review on state-of-the-art deep learning methods for process prediction,” *Artificial Intelligence Review*, vol. 55, no. 2, pp. 801-827, 2022.
- [4] M. T. Rasheed, D. Shi, “LSR: Lightening super-resolution deep network for low-light image enhancement,” *Neurocomputing*, vol. 505, pp. 263-275, 2022.
- [5] X. Y. Kong, L. Liu, Y. S. Qian, “Low-light image enhancement via poisson noise aware retinex model,” *IEEE Signal Processing Letters*, vol. 28, pp. 1540-1544, 2021.
- [6] S. Wang, D. M. Yang, R. Rong, X. Zhan, G. Xiao, “Pathology image analysis using segmentation deep learning algorithms,” *The American journal of pathology*, vol. 189, no. 9, pp. 1686-1698, 2019.
- [7] F. Huang, J. Zhang, C. Zhou, Y. Wang, J. Huang, L. Zhu, “A deep learning algorithm using a fully connected sparse autoencoder neural network for landslide susceptibility prediction,” *Landslides*, vol. 17, no. 1, pp. 217-229, 2020.
- [8] N. Sharma, R. Sharma, N. Jindal, “Machine learning and deep learning applications-a vision,” *Global Transitions Proceedings*, vol. 2, no. 1, pp. 24-28, 2021.
- [9] W. Zhang, H. Li, Y. Li, H. Liu, Y. Chen, X. Ding, “Application of deep learning algorithms in geotechnical engineering: a short critical review,” *Artificial Intelligence Review*, vol. 54, no. 8, pp. 5633-5673, 2021.
- [10] O. Aslan, A. A. Yilmaz, “A new malware classification framework based on deep learning algorithms,” *Ieee Access*, vol. 9, pp. 87936-87951, 2021.
- [11] O. Bamsile, A. Oluwasanmi, C. Ejayi, N. Yimen, S. Obiora, Q. Huang, “Comparison of machine learning and deep learning algorithms for hourly global/diffuse solar radiation predictions,” *International Journal of Energy Research*, vol. 46, no. 8, pp. 10052-10073, 2022.
- [12] W. Zhang, P. Zhuang, H. Sun, G. Li, S. Kwong, C. Li, “Underwater image enhancement via minimal color loss and locally adaptive contrast enhancement,” *IEEE Transactions on Image Processing*, vol. 31, pp. 3997-4010, 2022.
- [13] C. Ike, N. Muhammad. Separable property-based super-resolution of lousy image data. *Pattern Analysis and Applications*, 2020, 23: 1407-1420.
- [14] T. Zhang, T. Hou, S. Weng, F. Zou, H. Zhang, C. Chang, “Adaptive reversible data hiding with contrast enhancement based on multi-histogram modification,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 8, pp. 5041-5054, 2022.
- [15] K. Liu, X. Li, “De-hazing and enhancement method for underwater and low-light images,” *Multimedia Tools and Applications*, vol. 80, no. 13, pp. 19421-19439, 2021.
- [16] X. Sui, Z. He, G. Jin, D. Chu, L. Cao, “Band-limited double-phase method for enhancing image sharpness in complex modulated computer-generated holograms,” *Optics Express* vol. 29, no. 2, pp. 2597-2612, 2021.
- [17] R. Liu, Z. Jiang, S. Yang, X. Fan, “Twin adversarial contrastive learning for underwater image enhancement and beyond,” *IEEE Transactions on Image Processing*, vol. 31, pp. 4922-4936, 2022.
- [18] S. Wang, B. Kang, J. Ma, X. Zeng, M. Xiao, J. Guo, M. Cai, J. Yang, Y. Li, X. Meng, B. Xu, “A deep learning algorithm using CT images to screen for Corona Virus Disease (COVID-19),” *European radiology*, vol. 31, no. 8, pp. 6096-6104, 2021.
- [19] N. Sharma, R. Sharma, N. Jindal, “Machine learning and deep learning applications-a vision,” *Global Transitions Proceedings*, vol. 2, no. 1, pp. 24-28, 2021.
- [20] Y. R. Shrestha, V. Krishna, G. von Krogh, “Augmenting organizational decision-making with deep learning algorithms: Principles, promises, and challenges,” *Journal of Business Research*, vol. 123, pp. 588-603, 2021.
- [21] A. P. Pandian, “Performance evaluation and comparison using deep learning techniques in sentiment analysis,” *Journal of Soft Computing Paradigm (JSCP)*, vol. 3, no. 2, pp. 123-134, 2021.
- [22] Y. Y. Zhao, Q. J. Huang, “Image enhancement of robot welding seam based on wavelet transform and contrast guidance,” *Int. J. Innov. Comput. Inf. Control*, vol. 18, pp. 149-159, 2022.
- [23] Z. Zhao, B. Xiong, L. Wang, Q. Ou, L. Yu, F. Kuang, “Retinexdip: A unified deep framework for low-light image enhancement,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 3, pp. 1076-1088, 2021.
- [24] F. Rizani, “Image quality improvement using image processing method image brightness contrast and image sharpening,” *Multica Science and Technology (mst) Journal*, vol. 1, no. 1, pp. 6-12, 2021.
- [25] K. Zhang, K. Fang, W. Chen, F. Sun, Y. Song, L. Han, “Controlling the spreading of nanoliter-scale droplets on the fibers of fabrics for enhancing image quality and ink utilization,” *ACS Applied Materials & Interfaces*, vol. 13, no. 50, pp. 60581-60589, 2021.

# Fall Detection and Monitoring using Machine Learning: A Comparative Study

Shaima R.M Edeib<sup>1</sup>, Rudzidatul Akmam Dziyauddin<sup>2</sup>, Nur Izdihar Muhd Amir<sup>3</sup>

Computer System Engineering-Razak Faculty of Technology and Informatics (FTIR), Universiti Teknologi Malaysia, Kuala Lumpur, Malaysia<sup>1</sup>

Ubiquitous Broadband Access Networks Lab (U-BAN)-Razak Faculty of Technology and Informatics (FTIR), Universiti Teknologi Malaysia, Kuala Lumpur, Malaysia<sup>1,3</sup>

**Abstract**—The detection of falls has emerged as an important topic for the public to discuss because of the prevalence and severity of unintentional falls, particularly among the elderly. A Fall Detection System, known as an FDS, is a system that gathers data from wearable Internet-of-Things (IoT) device and classifies the outcomes to distinguish falls from other activities and call for prompt medical aid in the event of a fall. In this paper, we determine either fall or not fall using machine learning prior to our collected fall dataset from accelerometer sensor. From the acceleration data, the input features are extracted and deployed to supervised machine learning (ML) algorithms namely, Support Vector Machine (SVM), Decision Tree, and Naïve Bayes. The results show that the accuracy of fall detection reaches 95%, 97 % and 91% without any false alarms for the SVM, Decision Tree, and Naïve Bayes, respectively.

**Keywords**—Fall detection; machine learning; acceleration data; SVM; decision tree; Naïve Bayes; IoT

## I. INTRODUCTION

Human falling is feared because it may have both physical and psychological consequences. Compared to younger individuals, elderly have a higher chance of fall [1]. According to the World Health Organization (WHO), elderly represent 20 percent of the world's population [2]. By 2030, the global population aged 60 and more is estimated to reach 1.4 billion, and by 2050, it is estimated to increase from 962 million to 2.1 billion, compared to 2017 [1]. Falls has perturbing influence on the elderly, which may shorten their life expectancy. People older than 65 years often experience a fall every year at a rate of around one-third of the population. In addition to ageing, falling incidents also caused by a few other variables, including environment, level of physical activity, and cardiovascular problems. This can cause bodily harm, and the treatment for these injuries often requires a long stay in medical healthcare centers. The fear of falling, which limits older individuals' ability to engage in their Activities of Daily Life (ADL), is the major physiological problem they face. This concern leads to activity limitation, which may lead to insufficient gait balance and reduced muscle, both of which hinder an older adult's mobility and independence. Therefore, remote wearable technologies are necessary to monitor, detect, and avoid falls to enhance the quality of life in general (QoL). As a result of this, a knowledge of falls may be split into two categories: fall prevention and fall detection. It is possible to consider fall detection to be the process of detecting a fall via the use of sensors or cameras to contact medical personnel. For the

purpose of detecting and preventing falls, several systems that make use of a variety of sensors and algorithms have been created [2]. Referring to the dataset source [3], we learn that there is no machine learning applied, and SisFall dataset [4] are bias to western body structure, contras to this work preference that aims for Asian-based ADLs. For this reason, we have applied different machine learning algorithms to classify our previous collected data.

In this paper, we used the ASEAN experiments by own database, where do not depends on the other database and this is what distinguishes the work of this paper. Fig. 1 shows the overview of the system procedure.

Section II discusses the literature review. Section III overviews the machine learning algorithms employed in this work. The research methodology including data collection is explained in Section IV. Section V discusses the results. Section VI concludes this work.

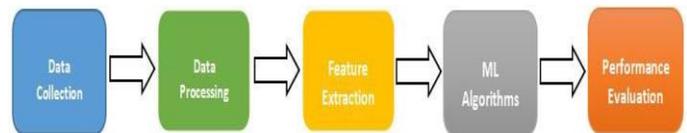


Fig. 1. The overview of system procedure.

## II. LITERATURE REVIEW

Ali et al. [5] compared the classification accuracy and execution speed of the J48 and to AdaBoost classifiers for fall detection. The proposed fall prototype was built by varying many distinguishing parameters, including velocity, geometric orientation, and geometric location. The results showed that 99.03 % fall detection accuracy. The execution time of the J48 classifier was 0.01 milliseconds, whereas the execution time of the AdaBoost classifier was 0.025 milliseconds.

Nevertheless, the effectiveness of these classifiers simplifies complex subjects, such as people wearing identical clothing or having the same backdrop colour. Min et al. [6] studied the area under the ROC curve (Receiver operating characteristic), they analysed the performance of faster recurrent neural network (RCNN). The proposed system identified falls by analyzing the situation. Based on deep learning and activity characteristics, they proposed a unique technique for recognizing human falls on furniture. Include other human characteristics, such as speed of motion, centroid, and aspect ratio. The outcome was an AUC of 0.941% and a

precision of 93%. Zhang et al. [7] developed a fall detector using the Support Vector Machine (SVM) algorithm. The detector was equipped with a single accelerometer worn at the waist. Accelerations in both directions, variations in acceleration, and other factors were among the features for machine learning. Their method successfully detected falls approximately 96.7% of overall cases. The suggested approach included the incorporation of an accelerometer into a mobile phone for the purpose of determining the occurrence of falls. The body-fixed sensor made detection more difficult, putting the mobile phone in a pocket or wearing it around the neck made it more difficult. About 93.3% of occurrences, the mobile phone system properly triggered the warning.

Using five wireless accelerometers and a wireless heart rate monitor, Tapia et al. [8] developed a real-time method for automated detection not only of physical activities, but also in certain situations, by utilizing a wireless heart rate monitor and five separate triaxial accelerometers. The shoulder, the wrist, the hip, the thigh, and the ankle were attached to the accelerometers. A predefined window size was used to recover the characteristics from the time and frequency domains of the signal. Some of these characteristics are the Fast Fourier Transform (FFT) peaks, variance, energy, and correlation coefficients. Classifiers, such as C4.5 and Naive Bayes were used in order to separate activity into the following three categories: postures (such as standing and sitting), activities (such as walking and cycling), and other activities (running, using stairs, etc.). When subject-dependent training was used, the recognition accuracy for these three classes was 94.6%, but subject-independent training resulted in just 56.3% of accuracy.

Xiong [9] also introduced a skeleton-based 3D consecutive-low-pooling neural network (S3D-CNN). Compared to existing methodologies, the proposed method fared the best on publicly available and user-collected datasets. Wang et al. [10] proposed a fall detection system comprised of many sensors. They used Multisource CNN Ensemble (MCNNE) architecture to enhance the accuracy of detection. They discovered that MCNNE outperforms both a single CNN structure and a multitude of ensembled bi-model structures. Hnoohom et al. [11] used sensor data from accelerometers and gyroscopes to compare the performance of conventional ensemble learning. Whether the sensor is placed on the arm or the waist, the study's results imply that strategies based on ensemble learning may improve detection accuracy.

Considering all the above aspects in the proposed system there is no analysis for the current data-set which used in this study. It shows the obtained results of comparison by using different machine learning algorithms to detect elderly fall.

### III. MACHINE LEARNING ALGORITHMS

ML gives the system the ability to learn from the dataset and the patterns in the data by using them as inputs. During the data gathering procedure, sensors offer information of several fall parameters. Then, machine learning techniques are employed to categorize or detect fall behaviour depending on the application requirements. The following is kinds of the machine learning (ML) algorithms that are commonly utilized for fall detection and prevention, and used in this work as

well [2].

#### A. Support Vector Machine

The support vector machine (SVM), which is a kind of supervised machine learning model, may be used to determine the location of a hyperplane in a space that has  $n$  dimensions (where  $N$  is the number of features that distinctly divide the data). Although the support vector machine, also known as an SVM, may be used for both, classification and regression analysis, the former is where its principal application resides. Linear and non-linear support vector machines (SVM) are the two different types of this kind of machine. The linear classifier works on the assumption that all of the data points may be linearly divided into groups. As a consequence of this, it differentiates between the two classes by picking the hyperplane that maximizes the margin in the best possible way. Before determining a discriminant function, the non-linear classifier that is most usually employed maps the data using a kernel. This step is followed by the determination of the function. This discriminant function is linked to the hyperplane in the space that has been transformed. In addition to this, the kernel is used for pattern analysis in a number of other machine learning techniques [2]. Support Vector Machine (SVM) has one of the highest fall classification accuracies among the machine learning methods examined [12]. Where, the findings revealed that the linear SVM was one of the optimal classifiers for this cross-dataset validation strategy, as it accurately discriminated a fall event from typical day-to-day activities with a great accuracy rate and comparably high sensitivity and specificity [13].

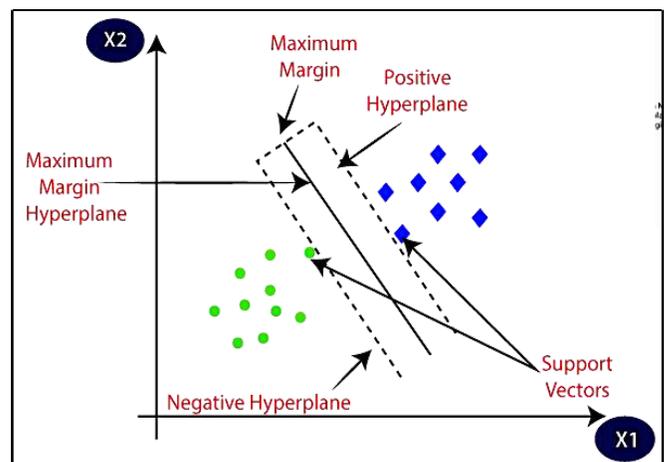


Fig. 2. Support vector machine.

Fig. 2 shows the classification of two distinct categories utilizing a decision boundary or hyperplane, which is this best decision boundary. Where, the aim of the SVM method is to generate the optimal line or decision boundary that divides  $n$ -dimensional space into classes, so that subsequent data points may be readily classified.

#### B. Decision Tree

A decision tree is a classifier that recursively splits the instance space. The decision tree consists of nodes that connect to form a rooted tree. This shows that the decision tree is a directed tree with a "root" node that lacks incoming edges.

Each internal node of a decision tree partitions the instance space into two or more sub-spaces according to a discrete function of the input attribute values. In the most common and straightforward example, each test examines a single attribute, splitting the instance space depending on the attribute's value. For quantitative attributes, the condition provides a range [3].

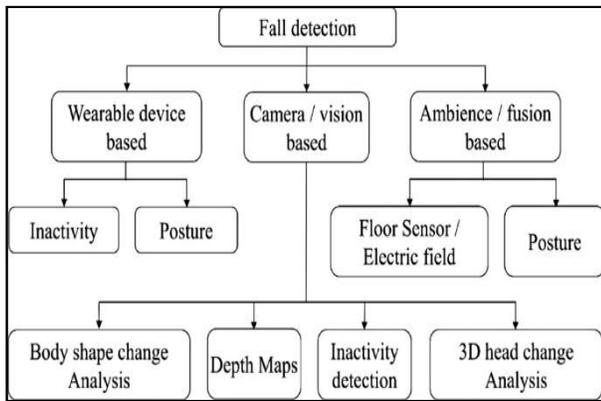


Fig. 3. Decision tree machine learning.

Fig. 3 illustrates different fall detection methods that make use of accelerometers, gyroscopes, or maybe both of them.

### C. Naive Bayes

The Naive Bayes algorithm is another supervised learning technique based on the Bayes Theorem. It is one of the simplest and most extensively used classification algorithms that may provide accurate predictions quickly. The Bayes theorem is used to generate classifications based on probability. On the basis of classes, uneven gait and falls may be immediately and readily identified [2]. It is a probabilistic classifier, which means that it makes its predictions based on the likelihood that an item would be found.

In Fig. 4 that has been shown below, it is an example that illustrates how Naive Bayes classifier has distinguished between the data points that have a fine border. The Gaussian curve in its original form has been applied here.

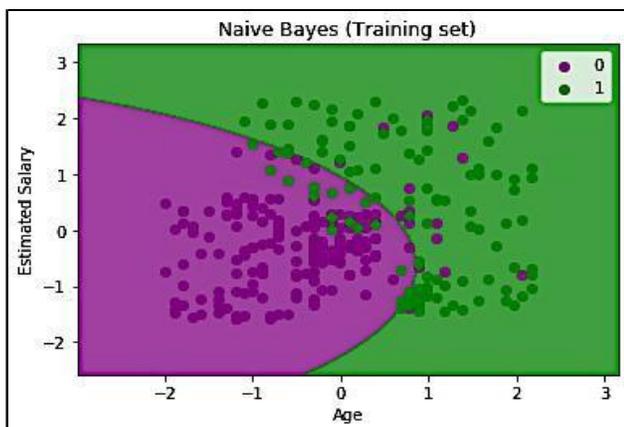


Fig. 4. Naive Bayes machine learning.

The preceding algorithms are generally used in fall

detection and prevention applications. Additionally, there are other algorithms such as Logistic Regression and Dynamic Time Wrapping, exist for similar applications (DTW). Another method for recognizing falls is to see them as an anomaly detection issue. Auto-encoders are utilized to detect falls in such systems. Auto-encoder learns features via ADL model training. Based on the reconstruction inaccuracy, fall actions are thus classified as an abnormality. It includes an encoder, a decoder, and a code layer. An encoder learns and compresses the input's essential characteristics. The code layer is the intermediate layer that includes important and compressed data information. In contrast, the decoder converts the data back into the original input. This approach may aid in reducing the complexity of data, obtaining required gait characteristics, and detecting unobserved falls [2].

## IV. METHODS

### A. Hardware Device Setup

The hardware involves a wearable device known as transmitter (FDS-Tx), which consist of the main controller (Arduino Pro mini), the wireless transmission module (XBee Pro) and the main component, a sensor (ADXL335 accelerometer). Before start the data collection, FDS-Tx will be attached to the volunteers' garment, specifically slightly above the right chest area. The device working system is, upon start, the user/volunteers' movements data will be recorded by accelerometer and sent to the receiver (FDS-Rx) via wireless transmission medium to the workstation. There, the computation take place to get the results of the volunteer's conditions (Fall or Normal). Details of the hardware description can be referred to work in [14]. Fig. 5 illustrates the setting of sensor on the user 0.

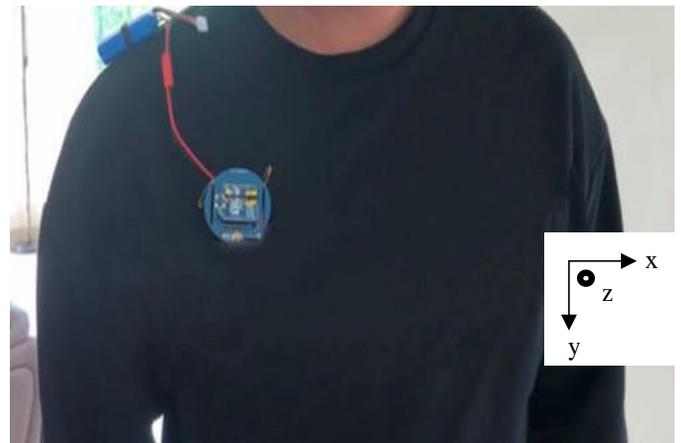


Fig. 5. Location of FDS-Tx on the user [26].

### B. Data Acquisition

ADXL335 three-axis accelerometer provides analogue voltage readings for X, Y, and Z acceleration (Fig. 6). An accelerometer can determine the tilt angle touching the earth by detecting the acceleration due to gravity. By measuring the dynamic acceleration, the accelerometer can determine the device's speed and direction of movement. Accelerometer with an analogue interface show accelerations by a range of voltage levels. In general, these values oscillate between the ground and the supply voltage. The micro-controller's ADC may then

be used to read this value. For detecting a fall using accelerometer, presently, there are two sorts of detection approaches: analytical methods and machine learning methods [15]. The X and Y axes have a bandwidth selection range of 0.5 Hz to 1600 Hz, while the Z axis has a bandwidth selection range of 0.5 Hz to 550 Hz. In general, accelerometer is low-power devices. Typically, the needed current is between a micro to milli amp [16].

The data acquisition is based on the Activity list as in the Table I. Three (3) activities are performed for 3 times each for better analysis process which will be discussed in the next section [3].

TABLE I. TEST SCRIPT

Code	Activity	Trial s	Duration (s)
F01	Fall forward while walking caused by a trip	3	15
F02	Fall to the right while walking caused by a trip	3	15
F03	Fall to the left while walking caused by a trip	3	15

Equipment setup are consisting of the FDS-Tx and FDS-Rx hardware, and safety mats in a closed venue with adequate sizes for 10 to15 seconds of straight walk. 10 volunteers are involved in this session with the detail of age, height, weight and gender are recorded for further analysis reference. During data acquisition, volunteers will walk straight on the normal carpeted composite structure, and then falls on a safety mats to reduce the forces during fall impact. Fig. 6 demonstrates the volunteer performing one of the activities.



Fig. 6. A volunteer performing F03 activity under researcher supervision [15].

In this part, we present an overview of the framework used to identify ADLs and falls occurrences, as well as an explanation of the activity recognition technique (Fig. 7).

C. Data Processing

Data implementation: Google Colab product used for machine learning applications was used to process the available data [12]. After the data was collected and the needed libraries and packages were obtained and imported, the process of implementing the coding started. As a precaution, a unique ID of the data-set was specified within the drive to allow a seamless download. The data on the status activities of the

elderly consists of four input features (X-axis, Y-axis, Z-axis accelerations, and total MAG) and 1 output feature (STATUS), which are viewed through the panda’s package.

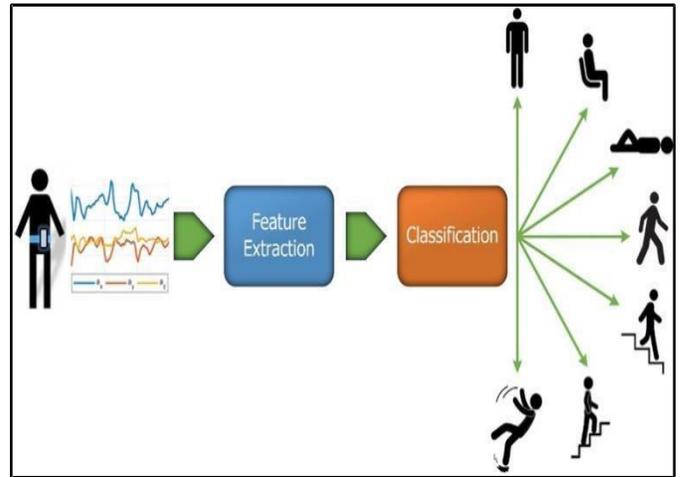


Fig. 7. Illustration of the activity recognition framework.

The data was distributed into Normal and Fall activities. Among 2245 activity, 1788 was Normal, and 457 was for Falling. Using Google Colab software, a categorical feature encoding was used. This feature converts non-numeric features to numbers for ease of machine learning. The STATUS of the Normal activity is precisely encoded as 1, while Fall activity is encoded as 0.

D. Data Set

Collected Data-set: The real datasets should be created, as current datasets including samples from 10 people that are physically different and with 3 different activities. Where number of trials for each activity is 3, and the total number of data is 84. Two (A01 and A02) out of ten participants only perform 2 activities because of personal health problem. To be clarified, the data set has been arranged as shown below in Fig. 8:

X-ACCEL	Y-ACCEL	Z-ACCEL	TOTAL MAG	STATUS
X-ACCEL is the acceleration of x-axis in g unit				
Y-ACCEL is the acceleration of y-axis in g unit				
Z-ACCEL is the acceleration of z-axis in g unit				
TOTAL MAG is the total magnitude of the three axis				
STATUS shows whether the person was fallen labeled as 'FALL' or was not fallen labeled as 'NORMAL'				

Fig. 8. Arranging of data set.

Accuracy measurement: To validate the accuracy of the results, the SVM machine learning algorithm was used and compared with the Decision Tree and Naive Bayes machine learning algorithms. The data is divided into 80-20% training-testing set splits where 1796 and 449 samples are used for training and testing, respectively.

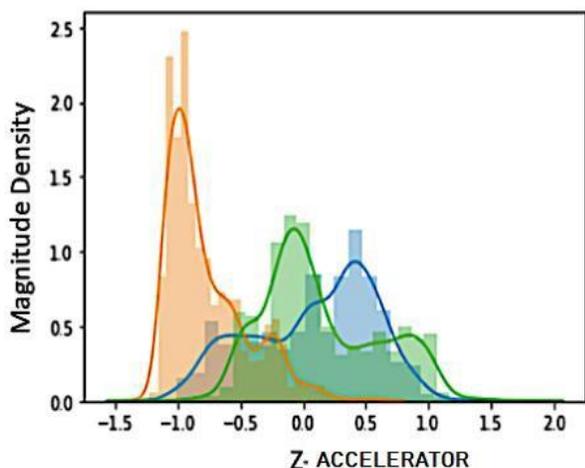


Fig. 9. Accelerometer input features.

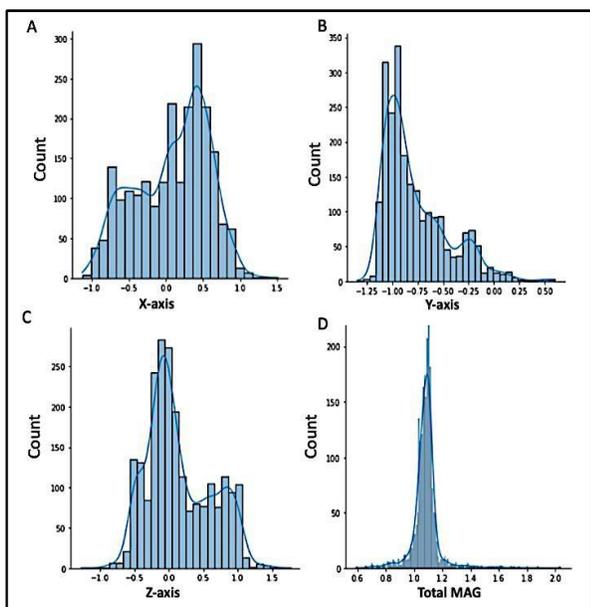


Fig. 10. Accelerometer input features of x-, y-, z- axes accelerations, and total MAG.

## V. RESULTS AND DISCUSSION

### A. Input Features Distribution

The data distribution of the X, Y, Z, and total MAG was assessed before applying the machine learning algorithms and found that the data was skewed and abnormally distributed (Fig. 9). A separate axial distribution of each axis was plotted and presented abnormal distribution (Fig. 10A, B, and C). However, the total mag feature exhibits a normal distribution (Fig. 10D). These results suggest that the dataset used in this study composes a suitable confusion matrix, which can be used to understand the classification model and correctly predict the possible errors.

### B. Machine Learning Accuracy Measurement

The training and testing accuracy of the proposed machine learning algorithms (SVM, Decision Tree, and Naive Bayes) was measured. However, the testing accuracy was used to rank

the algorithms as it offers consistent and more reliable results. Surprisingly, the Decision Tree algorithm provided the best accuracy (97%). SVM algorithm showed a relatively high accuracy of 95% as well (Fig. 11). Taken together, current measurements suggest that the best algorithm for fall detection is the Decision Tree machine learning algorithm. Indeed, employing Decision Tree will result in 100 % accuracy since a portion of training data is utilised for testing. The decision tree learns about the data during training, and if the same data is used to forecast today, it will provide the same outcome.

Therefore, a decision tree outperforms other machine learning algorithms.

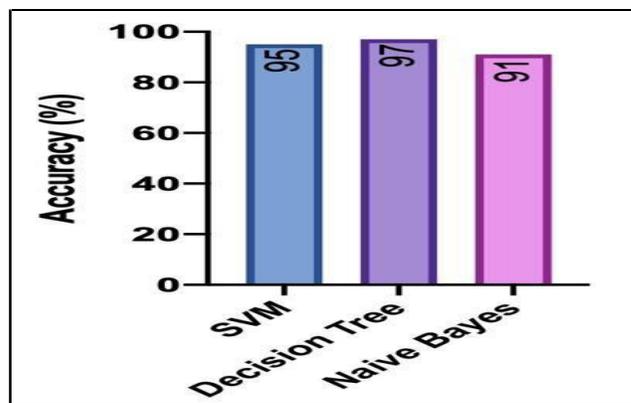


Fig. 11. Comparison of algorithms accuracy.

Fig. 12 shows the confusion matrix of the predicted results for Naive Bayes, SVM and decision tree algorithms. Where, DT carried out best performance comparative the other classification methods by getting 97% accuracy, whereas Naive Bayes shows the least performance by attaining 91% accuracy.

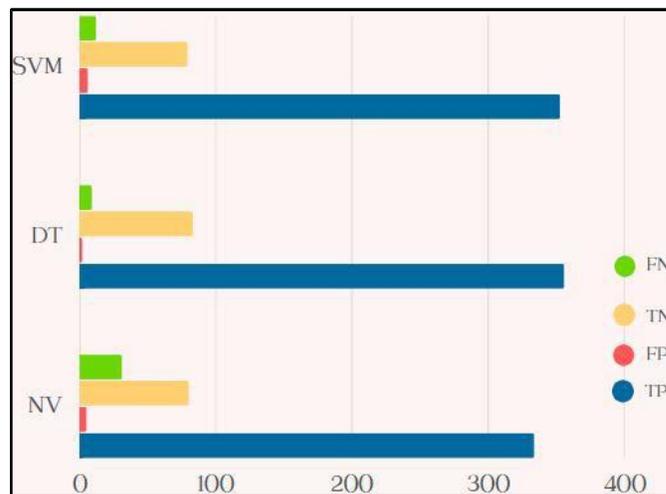


Fig. 12. Classification results using confusion matrix.

## VI. CONCLUSION

The purpose of this research is to compare some ML-based fall detection system as offline. It examines the systems based on a variety of characteristics including data-sets, confusion matrix and accuracy. The performance accuracy of the SVM,

Decision Tree, and Naive Bayes classification algorithms was tested using real-world acceleration data gathered from public databases. Using the training data, the internal parameters of these algorithms have been enhanced. Thereafter, the performance of the trained algorithms has been evaluated using the test data. The findings exposed that the SVM, Decision Tree, and Naive Bayes algorithms achieve an overall accuracy of 95%, 97%, and 91%, respectively. As next steps, it can be work on data generated by a combination of different types of sensors and vital signs sensors which may be worn by elderly people staying in old-age care homes or even their own homes. Also, the system may have another machine learning algorithms to support the end-to-end functionality.

#### ACKNOWLEDGMENT

This research work was funded by Universiti Teknologi Malaysia with the grant number R.K130000.7356.4B513 for financial support and device. Then, we would like to thank all those anonymous participants especially from UNITEN for their insightful contribution in the experiment.

#### REFERENCES

- [1] Komal Singh, Akshay Rajput, Sachin Sharma. "Human Fall Detection Using Machine Learning Methods: A Survey". International Journal of Mathematical, Engineering and Management Sciences. Vol. 5, No. 1, 161-180, 2020. (Received October 11, 2018; Accepted August 12, 2019) . <https://doi.org/10.33889/IJMEMS.2020.5.1.014>.
- [2] Sara Usmani, Abdul Saboor, Muhammad Haris, Muneeb A. Khan, Heemin Park. "Latest Research Trends in Fall Detection and Prevention Using Machine Learning: A Systematic Review". Sensors 2021, 21(15), 5134; <https://doi.org/10.3390/s21155134>. Received: 24 June 2021 / Revised: 16 July 2021 / Accepted: 24 July 2021 / Published: 29 July 2021.
- [3] Nur Izdihar Muhd Amir, Rudzidatul Akmam Dzuyaiddin, Norliza Mohamed, Nor Syahidatul Nadiah Ismail, Nor Saradatul Akmar Zulkifli, Norashidah Md Din. "Real-time Threshold-Based Fall Detection System Using Wearable IoT".
- [4] Sucerquia A, López JD, Vargas-Bonilla JF. SisFall: A fall and movement dataset. Sensors. 2017 Jan 20; 17(1):198.
- [5] Ali, S., Khan, R., Mahmood, A., Hassan, M., & Jeon, a. (2018). Using Temporal Covariance of Motion and Geometric Features via Boosting for Human Fall Detection. Sensors, 18(6), 1918. doi: 10.3390/s1806191.
- [6] Min, W., Cui, H., Rao, H., Li, Z., & Yao, L. (2018). Detection of the human Falls on Furniture Using Scene Analysis Based on Deep Learning and Activity Characteristics. IEEE Access, 6, 9324-9335. doi: 10.1109/access.2018.2795239.
- [7] Zhang, T., Wang, J., Liu, P., and Hou, J. Fall Detection by Wearable Sensor and One-Class SVM Algorithm. Lecture Notes in Control and Information Science, issue 345, pp. 858–863, 2006.
- [8] Tapia, E. M., Intille, S. S., Haskell, W., Larson, K., Wright, J., King, A., and Friedman, R. Real-Time Recognition of Physical Activities and Their Intensities Using Wireless Accelerometers and a Heart Rate Monitor. In Proceedings of the 11th IEEE International Symposium on Wearable Computers, pp. 37–40, 2007.
- [9] Xiong, X., Min, W., Zheng, W. et al. S3D-CNN: skeleton-based 3D consecutive-low-pooling neural network for fall detection. Appl Intell (2020). <https://doi.org/10.1007/s10489-020-01751-y>.
- [10] L. Wang, M. Peng and Q. Zhou, "Pre-Impact Fall Detection Based on Multisource CNN Ensemble," in IEEE Sensors Journal, vol. 20, no. 10, pp. 5442-5451, 15 May 15, 2020, doi: 10.1109/JSEN.2020.2970452.
- [11] N. Hnoohom, A. Jitpattanukul, P. Inluergsri, P. Wongbudsri and W. Ployput, "Multi-sensor-based fall detection and activity daily living classification by using ensemble learning", Proc. Int. ECTI Northern Sect. Conf. Electr. Electron. Comput. Telecommun. Eng. (ECTI-NCON), pp. 111-115, Feb. 2018.
- [12] Shing-Hong Liu, Wen-Chang Cheng. "Fall Detection with the Support Vector Machine during Scripted and Continuous Unscripted Activities". Sensors 2012, 12, 12301-12316; doi: 10.3390/s120912301.
- [13] Omar Aziz, Jochen Klenk, Lars Schwickert, Lorenzo Chiari, Clemens Becker, Edward J. Park, Greg Mori, Stephen N. Robinovitch. "Validation of accuracy of SVM-based fall detection system using real-world fall and non-fall datasets." Published: July 5, 2017. <https://doi.org/10.1371/journal.pone.0180318>.
- [14] Amir, Nur Izdihar Muhd, Rudzidatul Akmam Dzuyaiddin, Norliza Mohamed, Liza A. Latiff, and Nor Syahidatul Nadiah Ismail. "Development of Fall Detection Device Using Accelerometer Sensor." In 2021 IEEE 7th International Conference on Smart Instrumentation, Measurement and Applications (ICSIMA), pp. 93-98. IEEE, 2021.
- [15] Mohd Nadim, Aravindh S, Qayyum Khan, Rizwana. "An Automated Fall Detection System Using Accelerometer". International Journal of Scientific & Engineering Research Volume 9, Issue 7, July-2018 2033. ISSN 2229-5518. <https://www.ijser.org/researchpaper/An-Automated-Fall-Detection-System-Using-Accelerometer.pdf>.
- [16] Osama Zaid Salah, Sathish Kumar Selvaperumal, Raed Abdulla. "Accelerometer-based elderly fall detection system using edge artificial intelligence architecture". International Journal of Electrical and computer Engineering (IJECE). <https://www.researchgate.net/publication/360166901>.
- [17] Ali Chelli, Member, IEEE, and Matthias P'atzold, Senior Member, IEEE. "A Machine Learning Approach for Fall Detection and Daily Living Activity Recognition". Citation information: DOI 10.1109/ACCESS.2019.2906693, IEEE Access.
- [18] Marvi Waheed, Hammad Afzal, and Khawir Mehmood. "NT-FDS—A Noise Tolerant Fall Detection System Using Deep Learning on Wearable Devices". Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. <https://www.mdpi.com/journal/sensors>.
- [19] Nirmalya Thakur, and Chia Y. Han. "A Study of Fall Detection in Assisted Living: Identifying and Improving the Optimal Machine Learning Method". Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. <https://www.mdpi.com/journal/jsan>.
- [20] Sagar Chhetri, Abeer Alsadoon, Thair Al-Dala'in, P.W.C. Prasad, Tarik A. Rashid, Angelika Maag. "Deep Learning for Vision-Based Fall Detection System: Enhanced Optical Dynamic Flow". <https://arxiv.org/ftp/arxiv/papers/2104/2104.05744.pdf>.
- [21] Mitja Luštrek, Boštjan Kaluža. "Fall Detection and Activity Recognition with Machine Learning". Informatica 33 (2009) 205–212. [https://www.academia.edu/552485/Fall\\_detection\\_and\\_activity\\_recognition\\_with\\_machine\\_learning](https://www.academia.edu/552485/Fall_detection_and_activity_recognition_with_machine_learning).
- [22] Thiago B. Rodriguesa, Débora P. Salgadoa,b, Mauricio C. Cordeiroc, Katja M. Osterwald, Teodiano F. B. Filhod, Vicente F. de Lucena Jr.e, Eduardo L. M. Navesb, Niall Murray. "Fall Detection System by Machine Learning Framework for Public Health". The 8th International Conference on Current and Future Trends of Information and Communication Technologies in Healthcare (ICTH 2018).
- [23] Moiz Ahmed, BS, Nadeem Mehmood, PhD, Adnan Nadeem, PhD, corresponding author Amir Mehmood, MS, and Kashif Rizwan, MCS1. "Fall Detection System for the Elderly Based on the Classification.
- [24] Sadik Kamel Gharghan, Saleem Latteef Mohammed, Ali Al-Naji, Mahmood Jawad Abu-AlShaer, Haider Mahmood Jawad, Aqeel Mahmood Jawad, Javaan Chahl. "Accurate Fall Detection and Localization for Elderly People Based on Neural Network and Energy-Efficient Wireless Sensor Network". 23 October 2018. <https://pdfs.semanticscholar.org/e2c1/30e80e759fdd03b6c51aabb219a52b00d2fe.pdf>.
- [25] Sucerquia A, López JD, Vargas-Bonilla JF. SisFall: A fall and movement dataset. Sensors. 2017 Jan 20; 17(1):198.

# Online Teaching Design and Evaluation of Innovation and Entrepreneurship Courses in the Context of Education Internationalization

Chengshe Xing

Yiwu Industrial & Commercial College, School of Foreign Studies, Yiwu, 322000, China

**Abstract**—In the context of the internationalization of education nowadays, courses in innovation and entrepreneurship have been strongly promoted, and the content and number of topics, etc. of this type of courses are rapidly climbing. In order to enable target users to quickly select courses that they may be interested in, one changed collaborative filtering algorithm based on a multi-feature ranking model is used to extract and rank the features of online courses based on several factors, and then combine the collaborative filtering algorithm to recommend them to users. The results of experiment show that the numerical valuation of accuracy rate and recall rate of the improved algorithm are more than those of the other algorithm with different conditions, and in most cases higher than those of the LDA algorithm, and the user's evaluation of the recommendation effect also has the highest rating value of the improved algorithm, with the ratings of 4.3, 4.7 and 4.4 in the three groups, and the overall average score is 4.47, indicating that the improved algorithm has significant optimization performance and is suitable for teaching innovation and entrepreneurship in online courses.

**Keywords**—Online course; Collaborative filtering algorithm; Ranking model

## I INTRODUCTION

Traditional teaching suffers from single teaching mode, unbalanced teaching resources and low teaching efficiency. As the process of education informatization accelerates, more and more information technology and intelligent algorithms are applied to the education field and improve the existing teaching quality from them. With the increase of people's reliance on online teaching, educational resources have started to grow explosively in recent years, and many domestic and foreign experts have started to research on personalization of educational resources. The application of recommendation models to online teaching can not only improve the learning efficiency of learners and save their time in finding materials, but also improve the existing teaching models and educational resources. Based on this, recommendation models have become a hot research topic in the field of education. In recent years, innovation and entrepreneurship courses have been of high interest and attention, and in the context of the current internationalization of education, the content required for innovation and entrepreneurship courses has been increasing, which has led to an increasing demand for online courses in related directions [1]. The industry of teaching online courses has become increasingly mature, and course recommendations are often made in online courses to shorten the time of user

selection or to give suggestions for them [2]. For teaching recommendations in online courses, the core of the recommendation system is the recommendation algorithm used, and the performance of the recommendation algorithm directly affects the final recommendation results [3]. There are several recommendation methods, among which the most commonly used are different types of collaborative filtering algorithms. Collaborative filtering algorithms have the advantages of fast computational rate and high accuracy, but also have drawbacks such as cold start problem [4]. Therefore, in order to make the collaborative filtering algorithm fully applicable to the teaching recommendation of courses, the traditional collaborative filtering algorithm must be improved and achieve the desired effect. Currently, collaborative filtering is one of the most mature techniques in recommender systems, which uses the similarity of interests or features to find the nearest neighbors and then recommend the target users. Although this algorithm has high recommendation quality, it still faces the problem of data sparsity, where users generate more evaluations for mature domains and fewer evaluations for unfamiliar or new domains, thus causing the cold start phenomenon. If we can learn users' preferences or features from mature domains and use them in the recommendation of new domains, it will greatly alleviate the problem of data sparsity in new domains.

## II RELATED WORK

For the research and improvement of collaborative filtering algorithm, many experiments have been successfully conducted by scholars at home and abroad. Liu came up to a changed clustering-based collaborative filtering algorithm for reference by introducing a function of decay time and item attribute vector and characterizing items and user interest vector to describe users, and projecting recommendation candidate sets in clusters, and the result of experiment showed that the algorithm can solve the data sparsity and problem of new items [5]. Xu et al. established object meta-classification by introducing a new dependency function based on Gaussian kernel and extended classification method for information of input statistics data, and recommended results by computing the dependency between the features of classification set and the features set of target objects. The results of experiment show that, compared with conventional algorithms, the new hybrid has higher speed and better performance [6]. Chen's team calculates the similarity between users by combining the collaborative filtering algorithm with other algorithms, and computes users' calligraphy words in addition to the main

recommended calligraphy words based on the preliminary recommendation results to get the final recommendation results [7]. Panda team designed a collaborative filtering recommender based on normal filter for recommender systems to recommend personalized objects to strict users. The algorithm determines the average user rating for each object, computes the number of users who bought corresponding objects. Then uses min-max normalized way to find the number of users who have been normalized for each object in a specific range to scale the average user rating, and finally tested and found to predict user ratings more accurately [8]. Yu et al. came up to a cross-domain algorithm based on feature collaborative filtering construction and locally weighted linear regression by constructing features in different domains and using these features to represent different auxiliary domains. Also, they used a locally weighted linear regression model to solve the regression problem. Results of experiment show that this regression algorithm effectively solves the data sparsity problem by transferring useful data of knowledge from the auxiliary features domains [9].

Jiang's team proposed a slope-one algorithm to calculate the similarity between users by selecting trusted data and adding this similarity to the weighting factor of the changed algorithm to obtain the final recommendation equation. The base of new algorithm is the fusion of trusted data. User similarity under collaborative filtering algorithm acts more accurately than traditional algorithms [10]. Osval Montesinos-López et al. developed a project-based collaborative filtering package for multi-trait and multi-environment data and used it to study the prediction accuracy of precious data under phenotypic and genomic selection. The results of simulation experiment showed that package was more accurate for studying genomic prediction and data predictions are more accurately [11]. Zhang et al. designed a coverage-based collaborative filtering algorithm to provide brilliant recommendations for new users, improved previous collaborative filtering by reconstructing a decision class with detailed analysis of new user characteristics, and used a coverage-based the results showed that the improved algorithm significantly outperformed the existing working algorithm [12]. Li et al. in view of the increasingly serious problem of information overload and the fact that the traditional recommendation algorithm does not take the social relationship of users as the basis of recommendation, a combination algorithm with social information and dynamic time window is proposed. Through dynamic time window comparison, the time function is introduced to determine the corresponding time weight of user interest at different times. Finally, the practicability and effectiveness of the proposed method are verified. The experimental results show that the performance of the proposed algorithm is better than that of the traditional collaborative filter synthesis algorithm [13]. Yildirim studied the relevance between users through online social networks and used a multi-type improved collaborative filtering algorithm used for shopping recommendations, and the experimental results showed that the numerical valuation of accuracy rate under the changed algorithm recommendations is higher [14]. Han et al. based on the improved k-means clustering of small batches, an improved time weighted collaborative algorithm based on small batch

K-means is proposed. The algorithm combines Pearson correlation coefficient with k-means algorithm, uses the improved k-means clustering algorithm of small batch to cluster the sparse scoring matrix, and introduces Newton cooling time weighting to improve user similarity. The experimental results are obviously superior to the traditional algorithm in all aspects [15].

From the above research results, it can be found that there are a large number of studies related to the personalization technology of collaborative filtering algorithm, and a considerable number of studies in different fields are used to improve the traditional collaborative filtering algorithm, but there are relatively few studies on the personalized teaching technology based on algorithm about collaborative filtering, so the research is based on the personalization technology of algorithm about collaborative filtering to design an innovative entrepreneurial network for each target user. Therefore, the research is based on the personalization technology of algorithm about collaborative filtering to design an innovative entrepreneurship web course for each target user to enhance the learning interest and learning ability of target users in the context of internationalization of education.

### III IMPROVED COLLABORATIVE FILTERING ALGORITHMS BASED ON MULTI-FEATURE EXTRACTION RANKING MODEL AND ITS APPLICATION

#### A. Application of Collaborative Filtering Algorithm and Ranking Model to Online Courses

Along with the rapid development of internationalization of education, the content of innovation and entrepreneurship courses that can be learned, gradually increased, and at the same time the personalized requirements of course learning also increased. In this context, experiments are conducted to study the personalized recommendation of courses, and the collaborative filtering recommendation algorithm is widely used in the recommendation algorithm and is suitable for the context of internationalization of education under big data. There are two tasks in the collaborative filtering recommendation system, which are rating prediction and Top-N recommendation. The rating prediction is mainly to predict the rating of items not rated by users according to their characteristics, while the Top-N recommendation recommends the N most likely items of interest to users based on the rating and the rating prediction.

Collaborative filtering algorithms are classified into memory-based collaborative filtering algorithms, model-based collaborative filtering algorithms, and content-based collaborative filtering algorithms. Due to the current explosive growth in the amount of information on the Web, the rapid increase in the information available to users makes a single type of collaborative filtering algorithm no longer applicable, and practical applications also usually mix multiple types of methods to adapt to more complex practical situations [16]. While in Top-N recommendation mainly neighbor models, the most commonly used one is the K-nearest neighbor model. K-nearest neighbors use the K most similar items or user ratings for weighting and use them as a basis to predict user ratings for unknown items [17]. The nearest neighbor algorithm focuses on the similarity calculation, and the closer

the similarity result is 0 to 5, the higher the similarity is, and the closer the users or objects are to each other. The cosine similarity and Pearson coefficient for similarity calculation are shown in equation (1) and equation (2), respectively.

$$\omega \cos(y, z) = \frac{r_y \cdot r_z}{|r_y|^2 \cdot |r_z|^2} \quad (1)$$

$$\omega_{pearson}(y, z) = \frac{\sum_{b \in B(r_{y,b} - \bar{r}_y)(r_{z,b} - \bar{r}_z)} (r_{y,b} - \bar{r}_y)(r_{z,b} - \bar{r}_z)}{\sqrt{\sum_{b \in B(r_{y,b} - \bar{r}_y)^2} (r_{y,b} - \bar{r}_y)^2} \sqrt{\sum_{b \in B(r_{z,b} - \bar{r}_z)^2} (r_{z,b} - \bar{r}_z)^2}} \quad (2)$$

In Eq. (1) and (2),  $r_y$  and  $r_z$  are the target vectors of the neighboring targets  $y$  and  $z$ , respectively. Cosine similarity is suitable in the absence of ratings by and loss vector space pinch cosine values for similarity judging, while Pearson coefficient is suitable in the case of both ratings and can eliminate rating noise and minimize the influence of users differing in rating stringency [18]. After completing the similarity calculation, the aggregation method is used to predict the numbers of rating on unknown objects as shown in equations (3) and (4).

$$r_{c,s} = \frac{1}{\sum_{c' \in U} |sim(c, c')|} \sum_{c' \in U} sim(c, c') \cdot r_{c',s} \quad (3)$$

$$r_{c,s} = r_c + \frac{1}{\sum_{c' \in U} |sim(c, c')|} \sum_{c' \in U} sim(c, c') \cdot (r_{c',s} - r_{c'}) \quad (4)$$

Both  $r_c$  and  $r_{c'}$  in Eq. (3) and (4) are the average ratings of the target users  $c$  and  $c'$ , respectively. Eq. (3) represents the aggregation function that does not consider the difference in users' rating styles, while Eq. (4) corrects the aggregation function for the difference in rating styles by calculating the deviation between the weighted and used rating values and the corresponding mean score difference values of the target users. The similarity determination recommendation process is shown as Fig. 1.

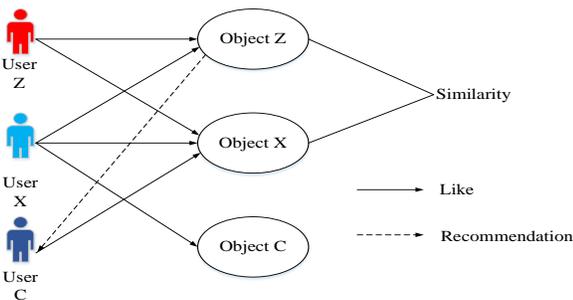


Fig. 1. User based collaborative filtering algorithm.

In practical applications, algorithms about collaborative filtering need to extract project features and target user features, and the extraction of user features is actually a supervised classification problem, so classification algorithms in machine learning can be used, including decision tree algorithms and linear classification algorithms, in addition to the nearest neighbor algorithm [19]. So, in order to be able to design a reasonable web-based distance course, the experiment requires feature extraction of course items and course learning users and reasonable recommendations using collaborative filtering algorithms, etc. The formula for the target user's number of rating of the item is shown in equation (5). In equation (5), the set of nearest neighbors is represented by  $S_u$ ,  $\bar{R}_u$  and  $\bar{R}_n$  mean the average ratings of users  $u$  and  $n$  on the objects each, the similarity on two users is represented by  $sim(u, n)$ , and the ratings of users  $n$  on the item  $j$  are represented by  $R_{n,j}$ .

$$P_{u,j} = \bar{R}_u + \frac{\sum_{n \in S_u} sim(u, n) \cdot (R_{n,j} - \bar{R}_n)}{\sum_{n \in S_u} sim(u, n)} \quad (5)$$

The content-based collaborative filtering algorithm will make the k most similar items (x1, x2, x3, x4 ....) based on the similarity of the target item x evaluated by target user .xk) get their corresponding similarity S (i.e., Sx1, Sx2, Sx3, Sx4 ..... Sxk), and then obtain a weighted average, which is based on the ratings of all those similar items by the target users, and then with the similarity of the target items as the weights, and use this result as the final rating of the target items by the target users. Finally, a number of top items are selected to be recommended to users based on the magnitude of the predicted value. A simple schematic of this algorithm is shown in Fig. 2.

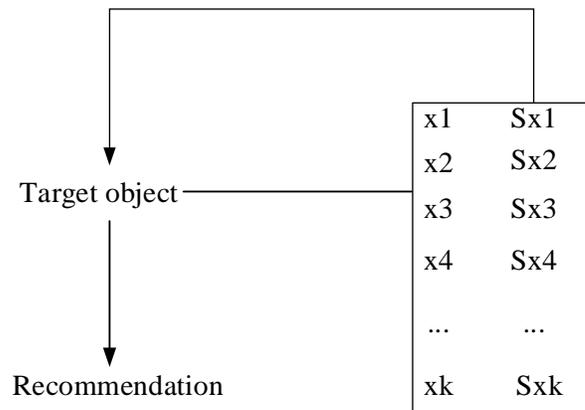


Fig. 2. Object based collaborative filtering algorithm.

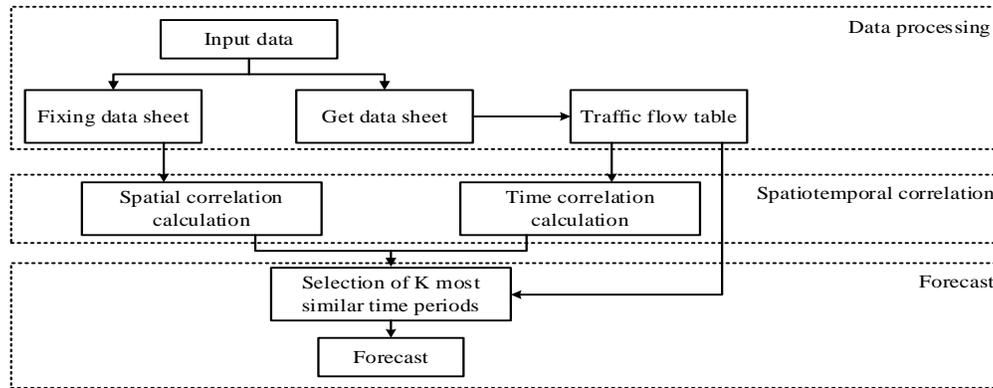


Fig. 3. Recommended schematic diagram of K proximity.

In order to fit the demands of the online teaching process for the main purpose, the dataset of the online course contains five data tables, and they are basic course information, course chapter information, basic user information, user learning and collection records and course creator information. Before the normal operation of the online course, the course information, course creator and user information need to be analyzed, at this time, the method of ranking learning can be used, and the ranking learning is integrated into the recommendation algorithm, and the weight parameters of multiple ranking models are learned by machine learning methods, and the best combination model is obtained in the training set to get better recommendation results [20]. Sorting learning usually requires the use of support vector machines, which are mainly divided into three types of methods: point-level methods, pair-level methods, and list-level methods, and their collaborative filtering k-neighborhood recommendation process is shown in Fig. 3.

The experiments are chosen to transform the ranking problem into a classification problem using a pairwise ranking method with fast training speed and moderate training complexity. Suppose from  $x_1$  to  $x_n$  are the feature vectors of documents, which are from  $d_1$  to  $d_n$ , at this point define a new training sample, where the positive samples are from  $x_1-x_2$  to  $x_1-x_n$ , negative samples are from  $x_2-x_1$  to  $x_n-x_1$ . Then a binary classifier is trained to classify these new samples. The document classification uses the support vector machine approach with a linear scoring function as shown in equation (6).

$$f(x) = \omega^t (x_u - x_v) \quad (6)$$

In equation (6)  $\omega$  is the marginal term and  $t$  is the time variable parameter. This method is integrated into the recommendation algorithm of course items, where the target users are ranked according to their preferences for the course items, and a matrix of users' access or learning time for the course items is established, and the input of the algorithm is a user-program evaluation matrix  $R(m, n)$  as shown in equation (7). The rows and columns in Eq. (7) represent the user and the item, respectively, and the element in the  $R_{ij}$ .

row of the  $j$  column in the matrix refers to the user's  $i$  numbers of rating on the object  $j$ .

$$R(m, n) = \begin{bmatrix} R_{1,1} & R_{1,2} & R_{1,3} & R_{1,4} & \cdots & R_{1,5} \\ R_{2,1} & R_{2,2} & R_{2,3} & R_{2,4} & \cdots & R_{2,5} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ R_{m,1} & R_{m,2} & R_{m,3} & R_{m,4} & \cdots & R_{m,5} \end{bmatrix} \quad (7)$$

### B. Improvement of Collaborative Filtering Algorithm Based on Multi-Feature Ranking Model

Conventional collaborative filtering algorithms based on ranking learning still have their limitations due to the shortcomings of traditional collaborative filtering algorithms such as cold start problem, sparsity problem and data density problem. Therefore, we designed an improved collaborative filtering algorithm based on multi-feature ranking and tested its applicability to the instructional design of an online course [21]. Considering each type of collaborative filtering algorithm, the experiment uses a user-based nearest neighbor recommendation algorithm. The similarity between all users is calculated, not by mixing all but by combining any two users. The similarity between the two is calculated by defining the collaborative filtering preference value of user  $x$  for course  $i$  as shown in equation (8). In equation (8),  $r(y, i)$  represents the rating of course 2 items by user  $y$  and  $i$  is the total number of users.

$$pref(x, i) = \frac{\sum_{y=1}^k sim(x, y) \cdot r(y, i)}{\sum_{y=1}^k sim(x, y)} \quad (8)$$

The obtained preference value results range from 1 to 5, and the higher the value is, means the higher the user's favorite preference for the course item. This preference value is used as a preference feature and applied in feature extraction, which is mainly associated with the calculation of similarity of the administrators, the time complexity of the calculation process and the number from users and items [22]. Since in the learning of online courses, courses with similar contents are not repeatedly studied by users, so in order to maximize the expression of users' interest bias, the

experiment uses a subject model for content preference feature extraction, i.e., a subject-based collaborative filtering algorithm for user preferences. Firstly, the course text information is abstracted into a subject vector, with the name and introduction as the course information, and the course learning record, personal filled-in interests and user learning record as the user profile. The course information and user profile are combined. The distribution of the courses that users have studied on the subject vector space is calculated as a weighted average, after which the similarity between the feature vector about users and the feature vector of courses is calculated [23]. After performing the similarity calculation, the user feature vectors and the subject-based user preferences are calculated as shown in Eqs. (9) and (10). In equation (9) and equation (10),  $\gamma_h$  represents the feature vector about the course  $h$ ,  $\gamma_g$  is the feature vector of the corresponding user,  $\gamma_h$  and  $\gamma_g$  are the same, and  $m$  is the number of courses that the corresponding user has taken.

$$\gamma_g = \frac{1}{m} \sum_{h=1}^m \gamma_h \tag{9}$$

$$pref(g, h) = \frac{\gamma_g \gamma_h}{\sqrt{(\gamma_g \gamma_g)(\gamma_h \gamma_h)}} \tag{10}$$

In addition to the above-mentioned features, feature extraction also takes into account the influence of the course lecturer itself and the popularity of the course. Since the influence of a course instructor cannot be directly quantified in the short term, the combined popularity of all the courses of the instructor is used as the influence of the instructor, and the popularity of the course should be studied first. The extraction of course popularity features should take something account number which is the numerical results of learners with user ratings, the number of scores with the ratio of the number of scores to the number of learners of the course. The feature extraction for calculating the popularity of a course is shown in equation (11).

$$val(u, i) = 0.2c_i + s_i \cdot d_i \tag{11}$$

In equation (11),  $c_i$  denotes the numerical valuation of learners for the course  $i$ ,  $d_i$  denotes the number of raters for the course  $i$ , and  $s_i$  denotes the all to aver numerical valuation after rating the course  $i$ . Since ratings are usually on a 5-point scale, to balance the any possible influence of the numerical results of learners and ratings on the course popularity value, the numerical result of learners is multiplied by a factor of 0.2. After deriving the absolute score of popularity, the calculation of similarity on popularity is shown in equation (12).

$$pref(u, i) = -\frac{r_i}{m} + 1 \tag{12}$$

In equation (12),  $m$  is the summary number of classes under the respective category, and  $r_i$  indicates the rank number on the course  $i$  in terms of the absolute value of its popularity among these courses. Similarly, the similarity value ranges from 0 to 1 and the closer to 1 means higher the similarity.

After obtaining the similarity of popularity, we can obtain the influence and popularity of the corresponding tutors as shown in equation (13). In equation (13),  $prefteach(u, i)$  represents the similarity of influence of tutors,  $prefhot(u, j)$  represents the similarity of popularity of different courses taught by tutors,  $m$  is the total number of courses, and  $C$  is a coefficient to consider the relationship between users and tutors, because users may have different values of  $C$  depending on whether they follow a tutor or not, and whether they have already taken a course. If you have not taken a lesson with a tutor and do not follow them, the value is 1.

$$prefteach(u, i) = C \cdot \frac{1}{m} \sum_{j=1}^m prefhot(u, j) \tag{13}$$

After all features are extracted and similarity is calculated, these recommendations are integrated using multi-feature ranking learning. A simple schematic representation of machine integration learning is shown in Fig. 4.

For a given user and project course, each pair of relationship is reflected as a vector  $x$ , each dimension in the vector which in this situation refers to the different features extracted, and the vector dimension is the number of features, at this time the ranking function is  $f(x) = Wx$  and  $W$  is the weight vector [24]. In this model, training sets are established according to the user learning of different courses, and the courses are grouped in each two items, and the first  $i$  group in the training set is shown in equation (14).

$$D_i = (x_{i1} - x_{i2}), i \in N^* \tag{14}$$

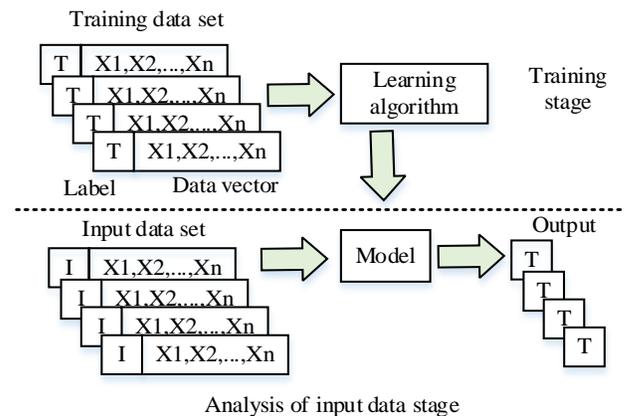


Fig. 4. Simple schematic diagram of machine sequencing calculation.

For each two sets of courses, the difference of the feature vectors of the two course items is used as the sample marker, and the samples that cannot be sorted by the sorting

relationship are ignored, and they are divided into two sets according to the different features of the two courses, which are positive and negative samples, after the sorting training is performed on these sample data [25]. The loss function for the ranking training is shown in equation (15). The same in equation (15)  $\omega$  is the marginal term and  $t$  is the time variable parameter.

$$lossf = \min \sum_{i=1}^m \max[1 - \omega^t (x_{i_1} - x_{i_2})] + \frac{1}{2} \|\omega\|^2 \quad (15)$$

The learning process of machine group sorting is then shown in Fig. 5. After that, the desired result of the predecessor is calculated.

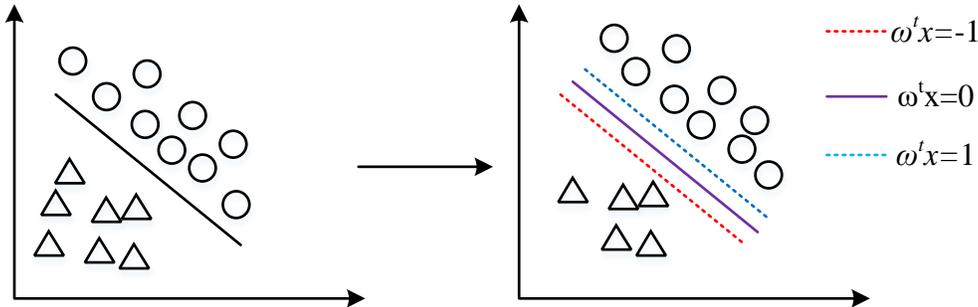


Fig. 5. Simple diagram of support vector machine learning.

#### IV DISCUSSION

The above research method optimizes the traditional collaborative filtering model and finally designs an improved collaborative filtering algorithm based on multi-feature ranking. In order to test the performance of the algorithm and explore whether the improved collaborative filtering algorithm is suitable for online teaching course recommendation, this section of the research content firstly selects a series of evaluation indexes for evaluating the strengths and weaknesses of different algorithms, and then compares the performance of evaluation indexes of different algorithms under the same test data set. The commonly used evaluation metrics for prediction scoring accuracy include mean absolute error, root mean square error, and standardized mean absolute error.

$$MAE = \frac{1}{I_T} \sum_{(u,i) \in I_T} |r_{ui} - \tilde{r}_{ui}| \quad (16)$$

Eq. (16) shows the formula of Mean Absolute Error (MAE) MAE, where  $I_T$  represents the test set.  $r_{ui}$  represents the actual rating of item  $i$  by user  $u$ .  $\tilde{r}_{ui}$  represents the predicted rating of item  $i$  by user  $u$ . The Mean Absolute Error can show the absolute error between the predicted and actual ratings, and the smaller the value, the better the recommendation effect of the algorithm.

$$RMSE = \sqrt{\frac{1}{I_T} \sum_{(u,i) \in I_T} (r_{ui} - \tilde{r}_{ui})^2} \quad (17)$$

Eq. (17) represents the calculation formula of root mean square error (RMSE) RMSE.

$$NMAE = \frac{MAE}{r_{\max} - r_{\min}} \quad (18)$$

Eq. (18) is the calculation formula of normalized mean absolute error (NMAE) NMAE.  $r_{\max}$  indicates the maximum value of the scoring range.  $r_{\min}$  indicates the minimum value of the scoring range. According to the above three average indicators, the commonly used recommendation algorithm and the improved algorithm in this paper are tested.

TABLE I. PERFORMANCE OF EVALUATION INDICATORS OF DIFFERENT ALGORITHMS IN THE SAME DATA SET

Recommended algorithm	MAE	NMAE	RMSE
Collaborative Filtering	26.54	13.27	5.15
Apriori	29.28	14.64	5.41
K-Means	18.35	9.18	4.28
BPNN	24.65	12.33	4.96
Algorithm in text	12.56	6.28	3.54

Table I shows five different types of recommendation algorithm models, including traditional collaborative filtering algorithm, Apriori algorithm based on association rules, K-Means algorithm based on clustering recommendation, Back Propagation Neural Network (BPNN) BPNN and improved collaborative filtering algorithm proposed in the paper. The MAE, NMAE and RMSE values of the five algorithms under the same recommendation data set are compared. According to Table I, the MAE values of the five algorithm models are 26.54, 29.28, 18.35, 24.65 and 12.56 respectively; NMAE values are 13.27, 14.64, 9.18, 12.33 and 6.28 respectively; The RMSE values are 5.15, 5.41, 4.28, 4.96 and 3.54 respectively. It can be seen that the improved collaborative filtering algorithm proposed by the research has good recommendation performance. Next, the algorithm is applied to the recommendation of online teaching resources for innovation and entrepreneurship courses, and the recommendation effect of collaborative filtering algorithm before and after improvement is compared. The accuracy

and recall rate are used to evaluate the recommendation accuracy of the recommendation model.

TABLE II. RELATIONSHIP BETWEEN RECOMMENDATION RESULTS AND USERS

User information		System recommendation	The system does not recommend
Users are interested		True-Positive ( $N_{tp}$ )	False-Negative ( $N_{fn}$ )
User is not interested		False-Positive ( $N_{fp}$ )	True-Negative ( $N_{tn}$ )

Table II shows the user's response to the recommendation system. As shown in Table II, there are four situations. That is, the user is interested in the recommended content of the recommendation system. The user is not interested in the recommended content of the recommendation system. The system does not recommend to the user but the user is interested. The system does not recommend to the user and is not interested.

$$P = \frac{N_{tp}}{N_{tp} + N_{fp}} \quad (19)$$

Eq. (19) shows the calculation formula of the recommended accuracy of the model. Accuracy can clearly represent the recommended performance of the model.

$$R = \frac{N_{tp}}{N_{tp} + N_{fn}} \quad (20)$$

Eq. (20) shows the calculation formula of the recommended recall rate of the model. The recall rate can indicate the probability that the content that users are interested in is recommended.

## V EXPERIMENTAL RESULTS AND COMPARATIVE ANALYSIS OF THE THREE ALGORITHMS

### A. Experimental Results of Feature Extraction and Model Training Prior

The experiment calculates user preferences based on common interval and topic-based user preferences, and randomly tries 80% of the training set and 20% of the training set. User course matrix test set. In order to calculate user preferences based on public filtering, use the marked value of KNN positive parameter K to capture the curve between the same value K, as shown in Fig. 6. Fig. 6 shows that the attack curve K is out of position. In order to fully study the needing rate, count and pick the value K 20 for the experiment.

Using Linear Discriminant Analysis (LDA) to calculate topic-based user preferences, the topic is the most important parameter in the algorithm, the experiment will be LDA as a separate recommendation algorithm for each user to select the corresponding topic vector as the most similar course as a result of the recommendation for that user, the number of topics and LDA recall results obtained are shown in Fig. 7. It can be seen that the recall rate gets one small increase with increase of the number of topics, and the number of

topics with a number of 100 is used for comprehensive consideration in Fig. 7.

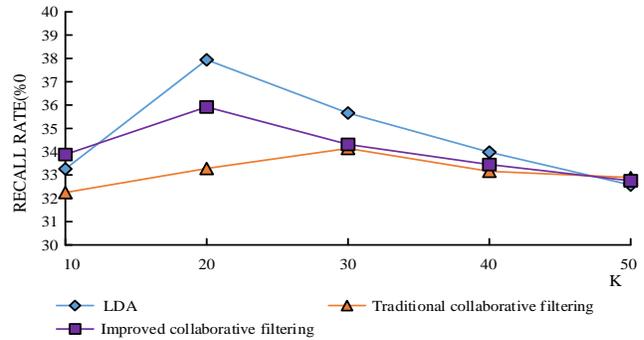


Fig. 6. Parameter K curve about recall rate.

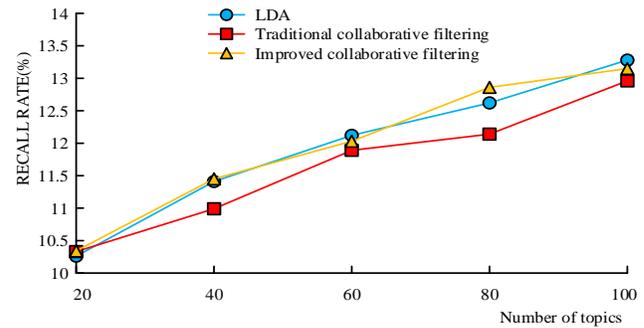


Fig. 7. Result curve of LDA about the number of topics.

### B. Analysis and Evaluation of the Results for the Online Course Algorithm

After learning the ranking function at the training, a ranking model is used to recommend courses according to their ranking and consequently generate the user's interest labels. The algorithm is then tested experimentally, along with two other algorithms, the regular algorithm about collaborative filtering and the linear discriminant analysis algorithm, and the effectiveness of the algorithms is measured and the results compared mainly by accuracy and recall.

The accuracy results of the three algorithms with different number of recommendations are displayed in Fig. 8. It can be seen, which are the truth that the accuracy rates of the three algorithms show a significant upward trend with the increase in the number of recommendations, which means that the recommendation algorithms should increase the number of recommendations from Fig. 8. When the number of recommendations is small, the accuracy rate of the improved algorithm is significantly higher than that of the traditional algorithm and the LDA algorithm, and when the number of recommendations gradually increases, the accuracy rate of the improved algorithm is still significantly higher than that of the traditional collaborative filtering, and at this time, although the accuracy rate is higher than that of the LDA algorithm, it will no longer be significant as the number of recommendations increases, indicating that algorithm which came up from the research is significantly better than the traditional algorithm and better than the LDA in general.

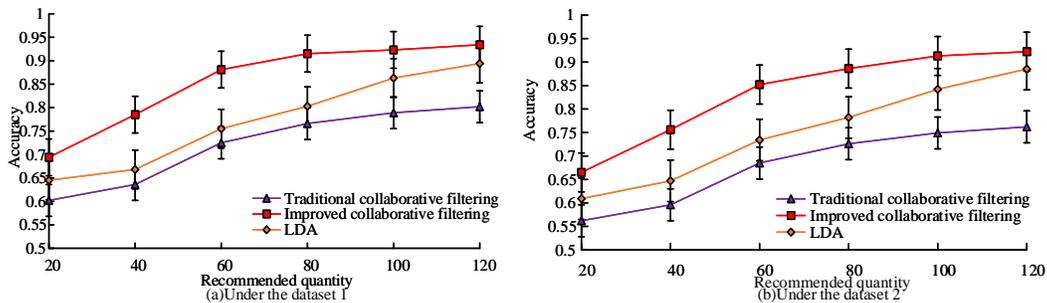


Fig. 8. Recommended quantity and accuracy results of three algorithms.

Fig. 9 shows the results of canceling many of the three proposed algorithms. As shown in Fig. 9, the cancellation rate of these three algorithms also increases with the increase of the number of recommendations. Among the three algorithms, the improved filter synthesis algorithm always has higher inverse speed than the traditional algorithm. When the recommendation number is low, the inverse value of LDA Algorithm is lower than the improved algorithm. The difference between LDA output coefficient and standard algorithm output coefficient gradually decreases, and the amount of recommendation increases, slightly higher than the improvement rate of the algorithm; among them, 120 recommendations.

For the same number of comments, the impact of the number of consumer training on the repeatability results is shown in Fig.10. Fig. 10 shows that the download rate of the three algorithms usually increases with the increase of the number of user training. When the number of user training is more, the calculation can play a better role. The extraction rate of the improved link filtering algorithm between the three algorithms is much higher than that of the traditional

algorithm, and is always higher than that of the LDA Algorithm. When the number of training users increase, the download rate of the improved algorithm is significantly higher than the LD rate. The table improved algorithm is better than the other two algorithms, and when the number of courses increases in a certain amount of pasta, the advantage of the improved algorithm is greater.

The impact of the number of user learning courses on the accuracy results is shown in Fig. 11. From Fig. 11, it can be seen that the accuracy of the three algorithms shows an overall upward trend as the number of user-learning courses increases, indicating that the algorithms can play a better recommendation effect when the number of user-learning courses is more. The accuracy of the improved algorithm among the three algorithms is always higher than the other two algorithms, and the difference between the improved algorithm and the LDA is more significant as the number of courses increases, indicating that the improved algorithm will have a superior performance as the number of courses increases.

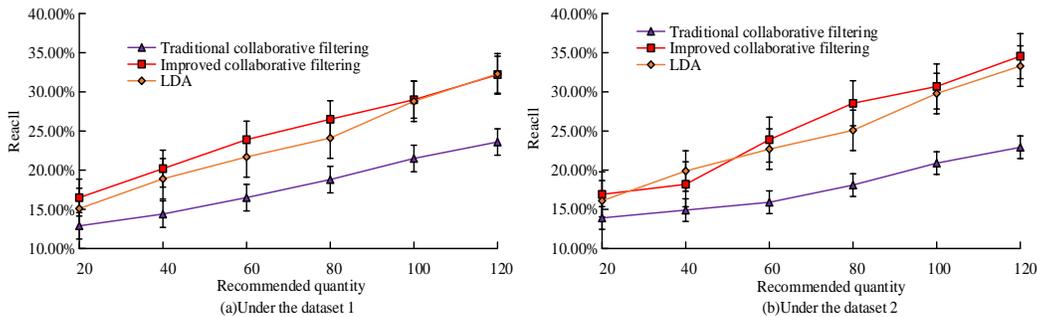


Fig. 9. Results of recommended quantity and recall rate of three algorithms.

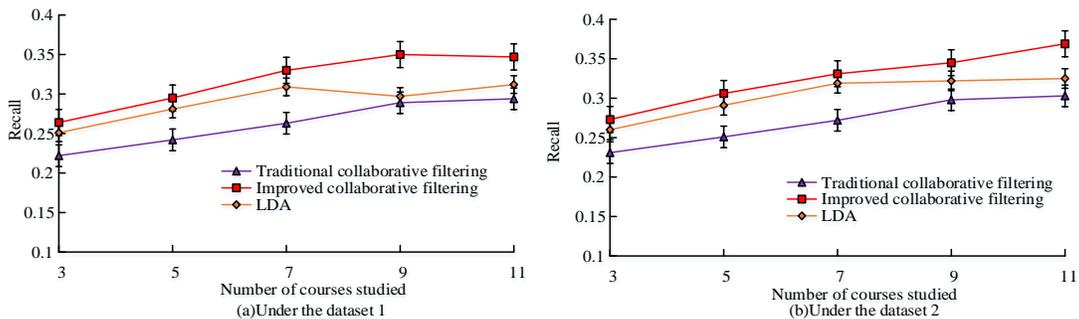


Fig. 10. Results of number of learning courses and recall rate of three algorithms.

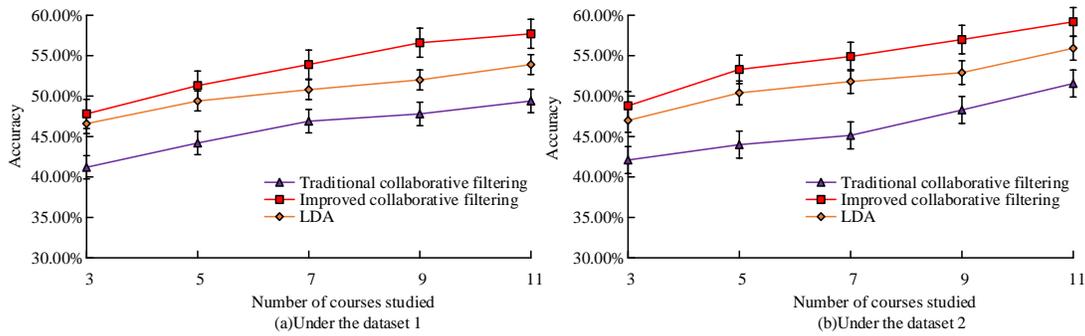


Fig. 11. Results of number of learning courses and accuracy of three algorithms.

The results of the recommendations were transmitted to the target users, the target users were divided into two large flat-rate groups, the results of the evaluation of the target users were taken from the sample and the results of each algorithm assessment were broken down into an average of five groups and then the average result for each group was calculated, - obtain the results of the evaluation referred to

in Fig. 12. Fig. 12 shows that, with the exception of the third group, the improved algorithm rates are higher than the other two and that the difference between the improved algorithm and the third group algorithm is lower, which fully indicates that the improved algorithm designs a course with higher performance than the class and is suitable for online courses training.

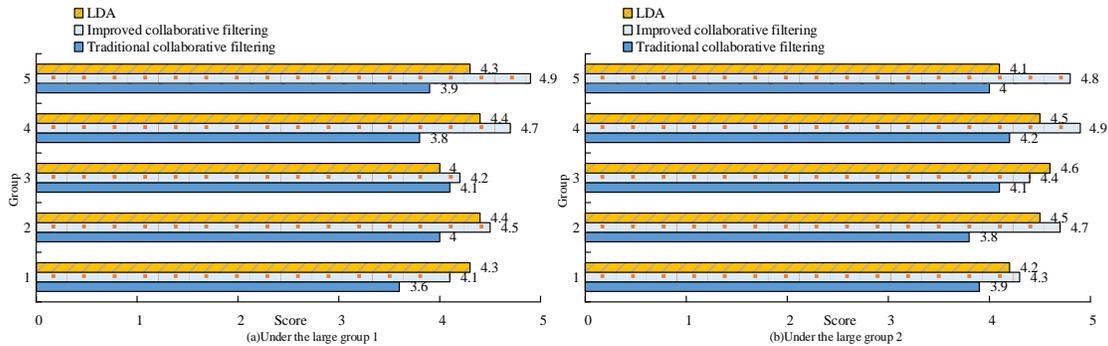


Fig. 12. Comparison of evaluation results of three algorithms.

## VI CONCLUSION

The research uses a modified cooperation filtering algorithm based on a multifunctional classification model, which is tested and compared with the regular cooperation filtering algorithm and LDA algorithm after sample training. Experimental results show that the maximum accuracy and recall rate of the improved algorithm is 92.2% and 32.2% respectively for different recommended quantities which are significantly higher than the traditional algorithm; the maximum level of accuracy with an improved algorithm cancellation rate is 57.7% and 34.7% respectively for different numbers of learning courses which are significantly higher than the other two algorithms. The overall mean of the improved algorithm is 4.46, which is higher than the other two algorithms in 5 with 5. Experimental results show that the improved algorithm works much better than the traditional algorithm and, in many cases, better or much better than the LDA algorithm. Although some results have been obtained in the study, the total number of samples selected in the experiments is still small and the overall shortage of random sampling results due to the small number of samples is the main direction for further studies and optimization in the future.

## VII FUTURE WORK

In this study, an improved collaborative filtering algorithm based on a multi-feature ranking model is used to construct a recommendation model for teaching innovation and entrepreneurship online courses in the context of internationalization of education. After testing the performance of the model, it is found that the constructed model has better recommendation effect, higher accuracy rate, and higher user satisfaction for innovation and entrepreneurship online courses. Although the research conducted in this paper has achieved certain results and optimized the traditional collaborative filtering algorithm to a certain extent, the following shortcomings still exist.

- 1) Although the improved collaborative filtering algorithm is applied to the course recommendation model, it may cause some bias to the experimental results because more pre-processing is not performed on the adopted educational data set.
- 2) Whether the constructed recommendation algorithm model is applicable to other e-learning recommendations for different majors is not explained in the paper. Subsequent research should attempt to apply the model to more online course recommendations for different majors.

3) In the current recommendation field, there are also more studies on the optimization of collaborative filtering algorithm, and the subsequent research should combine more advanced algorithms to optimize collaborative filtering and thus improve the recommendation accuracy.

#### ACKNOWLEDGEMENT

The research is supported by: 2022 Key research project of Yiwu Industrial & Commercial College: Teaching Evaluation Research in Higher Institutions against Education Internationalization (ZD2022WY322-01).

#### REFERENCES

- [1] N. Lin, Y. Lin, "Research on the Integration of Innovation and Entrepreneurship and Ideological and Political Courses in Universities under the Background of Internet Era", *Journal of Physics: Conference Series*, 2021, vol. 1852(4), pp. 42033-42037, 2021.
- [2] J. Lee, L. Martin, "Investigating Students' Perceptions of Motivating Factors of Online Class Discussions", *International Review of Research in Open & Distance Learning*, 2017, vol. 18(5), pp. 1492-1496, 2017.
- [3] W. Zhang, X. Zhou, W. Yuan, "Collaborative Filtering Algorithm Based on Improved Time Function and User Similarity", *Journal of Physics: Conference Series*, vol. 1757(1), pp. 12080-12087, 2021.
- [4] D. F. Meng, N. Liu, M. X. Li, et al. "An Improved Dynamic Collaborative Filtering Algorithm Based on LDA", *IEEE Access*, vol. 9(2), pp. 32-37, 2021.
- [5] X. Liu, "An improved clustering-based collaborative filtering recommendation algorithm" *Cluster Computing*, vol. 20(2), pp. 1281-1288, 2017.
- [6] L. B. Xu, X. S. Li, Y. "Guo Gauss-core extension dependent prediction algorithm for collaborative filtering recommendation", *Cluster Computing*, vol. 22(4), pp. 11501-11511, 2019.
- [7] S. Chen, L. Huang, Z. Lei, et al. "Research on personalized recommendation hybrid algorithm for interactive experience equipment", *Computational Intelligence*, vol. 36(3), pp. 1348-1373, 2020.
- [8] S. K. Panda, S. K. Bhoi, M. Singh, "A collaborative filtering recommendation algorithm based on normalization approach", *Journal of Ambient Intelligence and Humanized Computing*, vol. 11(1), pp. 4643-4665, 2020.
- [9] X. Yu, Q. Peng, L. Xu, et al. "A selective ensemble learning based two-sided cross-domain collaborative filtering algorithm" *Information Processing & Management*, 2021, vol. 58(6), pp. 102691-102702, 2021.
- [10] L. Jiang, Y. Cheng, L. Yang, et al. "A trust-based collaborative filtering algorithm for E-commerce recommendation system", *Journal of ambient intelligence and humanized computing*, vol. 10(8), pp. 3023-3034, 2019.
- [11] A. Osval, "Montesinos-López, Francisco Javier Luna-Vázquez, Abelardo Montesinos-López, et al. An R Package for Multitrait and Multienvironment Data with the Item-Based Collaborative Filtering Algorithm", *The Plant Genome*, vol. 11(3), pp. 1-16, 2018.
- [12] Z. Zhang, Y. Kudo, T. Murai, et al. "Improved covering-based collaborative filtering for new users' personalized recommendations", *Knowledge and Information Systems*, vol. 62(5), pp. 3133-3154, 2020.
- [13] D. Li, C. Wang, L. Li, et al. "Collaborative filtering algorithm with social information and dynamic time windows", vol. 52(3), pp. 261-272, 2021.
- [14] S. Z. M. Yildirim, "A new similarity coefficient for a collaborative filtering algorithm", *Communications Faculty of Science University of Ankara*, vol. 59(2), pp. 41-54, 2017.
- [15] X. Han, Z. Wang, H. J. Xu, "Time-Weighted Collaborative Filtering Algorithm Based on Improved Mini Batch K-Means Clustering", *Advances in Science and Technology*, vol. 1059(1), pp. 309-317, 2021.
- [16] N. Liu, M. X. Li, H. Y. Qiu, et al. "A hybrid user-based collaborative filtering algorithm with topic model", *Applied Intelligence*, vol. 12(51), pp. 7946-7959, 2021.
- [17] C. Ajaegbu, "An optimized item-based collaborative filtering algorithm", *Journal of Ambient Intelligence and Humanized Computing*, vol. 51(1), pp. 5261-5272, 2021.
- [18] H. Wang, Z. Shen, S. Jiang, et al. "User-based Collaborative Filtering Algorithm Design and Implementation", *Journal of Physics: Conference Series*, vol. 1757(1), pp. 012168-012173, 2021.
- [19] Y. Fan, H. Ma, Z. Chen, et al. "Research and Application of Algorithm Based on Maximum Expectation and Collaborative Filtering in Recommended System", *Journal of Physics Conference Series*, vol. 1754(1), pp. 012205-012209, 2021.
- [20] A. A. Alwan, H. Ibrahim, N. I. Udzir, "A Model for Ranking and Selecting Integrity Tests in a Distributed Database", *International Journal of Information Technology & Web Engineering*, vol. 5(3), pp. 65-84, 2017.
- [21] Y. Yang, H. Yao, R. Li, et al. "A collaborative filtering recommendation algorithm based on user clustering with preference types", *Journal of Physics: Conference Series*, vol. 1848(1), pp. 012043-012049, 2021.
- [22] Wang L. Application Model of Collaborative Filtering Algorithm Recommended for Pollution Source Information [J]. *Software Engineering and Applications*, 2020, 09(5):345-351.
- [23] R. Jenke, A. Peer, M. Buss, "Feature Extraction and Selection for Emotion Recognition from EEG", *IEEE Transactions on Affective Computing*, vol. 5(3), pp. 327-339, 2017.
- [24] K. L. Mak, K. Yiu, Z. Feng, "Feature extraction of the patterned textile with deformations via optimal control theory", *Discrete and Continuous Dynamical Systems - Series B (DCDS-B)*, vol. 16(4), pp. 1055-1069, 2017.
- [25] M. E. Banihabib, M. H. "Shabestari Fuzzy Hybrid MCDM Model for Ranking the Agricultural Water Demand Management Strategies in Arid Areas", *Water Resources Management*, vol. 31(1), pp. 495-513, 2017.

# Investigating Group Distributionally Robust Optimization for Deep Imbalanced Learning: A Case Study of Binary Tabular Data Classification

Ismail. B. Mustapha<sup>1</sup>, Shafaatunnur Hasan<sup>2</sup>, Hatem S Y Nabbus<sup>3</sup>, Mohamed Mostafa Ali Montaser<sup>4</sup>, Sunday Olusanya Olatunji<sup>5</sup>, Siti Maryam Shamsuddin<sup>6</sup>

Computer Science Department-School of Computing, Universiti Teknologi Malaysia, 81310 Skudai, Johor, Malaysia<sup>1, 2, 3, 4, 6</sup>

Higher Institute for Science and Technology, Al\_shomokh – Tripoli, Libya<sup>4</sup>

Department of Computer Science-College of Computer Science and Information Technology, Imam Abdulrahman Bin Faisal University, Dammam, Saudi Arabia<sup>5</sup>

**Abstract**—One of the most studied machine learning challenges that recent studies have shown the susceptibility of deep neural networks to is the class imbalance problem. While concerted research efforts in this direction have been notable in recent years, findings have shown that the canonical learning objective, empirical risk minimization (ERM), is unable to achieve optimal imbalance learning in deep neural networks given its bias to the majority class. An alternative learning objective, group distributionally robust optimization (gDRO), is investigated in this study for imbalance learning, focusing on tabular imbalanced data as against image data that has dominated deep imbalance learning research. Contrary to minimizing average per instance loss as in ERM, gDRO seeks to minimize the worst group loss over the training data. Experimental findings in comparison with ERM and classical imbalance methods using four popularly used evaluation metrics in imbalance learning across several benchmark imbalance binary tabular data of varying imbalance ratios reveal impressive performance of gDRO, outperforming other compared methods in terms of g-mean and roc-auc.

**Keywords**—Class imbalance; deep neural networks; tabular data; empirical risk minimization; group distributionally robust optimization

## I. INTRODUCTION

Owing to increased data availability, novel learning architectures and accessibility to commodity computational hardware devices, deep neural networks (DNNs) have become the de facto tool for a wide range of machine learning (ML) tasks in recent times; leading to state-of-the-art performance in several computer vision, natural language processing and speech recognition tasks. DNNs are characterized by several layers of hidden units that enable learning of useful representations of a given data for improved model performance [1, 2]. This alleviates the need for domain experts and hand-engineered features, a common prerequisite for traditional ML methods.

A pervasive problem that has plagued traditional ML methods in the last couple of decades which DNNs are not immune to is the class imbalance problem [3-6]. This problem, also termed long-tailed data distribution problem in computer vision, occurs when the distribution of the constituent classes

of a training data is highly disproportionate such that one or more classes have significantly larger number of training samples (majority class(es)) than other(s) (minority class(es)). Given that most ML methods are built to minimize the overall classification error with the assumption that each sample contributes equally, the learning algorithms tend to be bias towards the majority class; thus, resulting in partial or total disregard of the discriminative information of the minority classes by the learning algorithm. What makes this problem even more interesting is that, in most cases, the minority classes are often the classes of interest. Several manifestations of this problem abound in many real-life application domains of ML like medical diagnosis [7, 8], fraud detection [9-11], flight delay prediction [12, 13] amongst others.

The knowledge that learning from imbalance data negatively impacts the performance of DNN has resulted a marked increase in research on deep learning-based approaches to tackle the problem in recent years, with findings showing that traditional approaches to addressing imbalance problems can be successfully extended to DNN [3, 5]. Thus, many deep learning studies have addressed the imbalance problem at the data level mainly by data resampling [14-16] while some have done so at the classifier level, largely through cost sensitive learning approaches [17-22]. Oversampling and undersampling are two common data resampling approaches used in DNN. However, the susceptibility of the former to noise and overfitting due to added samples [23] as well as the characteristic loss of valuable information peculiar with the latter [3] remain major drawbacks of this category of imbalance methods. On the other hand, the core idea behind the cost sensitive methods is to assign different misclassification cost/weights to the training samples to scale up/down the misclassification errors depending on the class they belong [17, 24]. While there are several implementations of this method, the most commonly used cost sensitive approach in imbalanced deep learning research is reweighting [20, 25], where weights are assigned to different class samples based on either the inverse of the class frequencies [20, 26, 27] or their square root [28]. Despite its widespread adoption in DNN, reweighting methods have been found to be unstable in severely imbalanced cases; yielding poor performance that compromise the performance of the majority class [20, 23, 29,

30]. Inspired by the drawbacks of the commonly used methods, this study seeks to address the imbalance problem from a different perspective, through the learning objective.

The canonical training objective in DNN is the empirical risk minimization (ERM) which entails minimizing the average per sample training loss over the entire training data [31]. This training objective has the capacity to fit a given training data perfectly and still produce impressive accuracy on an unseen test data [32]. However, training a DNN using such objective on an imbalanced data has been shown to be bias to the majority class samples despite fitting the training data perfectly in most cases [32, 33]. The trained DNN is unable to generalize the learnt representations to the minority class samples at inference/test phase. In contrast to ERM, this study explores minimizing the maximum between the majority and minority class losses for improved imbalance learning. This is analogous to distributionally robust optimization (DRO) which seeks to minimize the expected loss over possible test distributions that the model is expected to perform well on [34, 35]. Specifically, group DRO (gDRO) proposed in [32] is investigated in this study in the context of classical class imbalance problem in DNN where the training and test data are similarly imbalanced. Rather than seeking reduction in the generalization gap between the training and test accuracies of the worst group, the performance of gDRO on binary imbalance datasets of varying imbalance ratios is investigated in comparison to popular imbalance methods in DNN. The performance of these methods is compared using four popular evaluation metrics in imbalance learning.

The rest of this article is outlined as follows: Section II provides the requisite background on ERM and gDRO followed by Section III which contains the methodology as it relates to the benchmark datasets, selected imbalance methods, DNN architecture and other experimental settings. The experimental results and discussions are presented and discussed in Section IV and Section V respectively before concluding in Section VI.

## II. THEORETICAL BACKGROUND ON ERM AND GDRO

Given a training data  $\mathcal{D}_t = \{(x_i, y_i)\}_{i=1}^N$ , where  $y \in \mathcal{Y}$  and  $x \in \mathcal{X}$  represent the target labels and the input features respectively,  $f_\theta$  is a prediction function parameterized by  $\theta$  that learns to correctly map each input feature  $x_i$  to the corresponding output label  $y_i$ . The aim is to find the set of parameters  $\theta$  that minimize the risk in (1).

$$\min_{\theta} \mathbb{E}_{(x,y) \sim \mathcal{D}_t} [\ell(y, f_\theta(x))] \quad (1)$$

where  $\ell$  and  $f_\theta(\cdot)$  stand for the loss function and predicted output respectively. Equation (1) is approximated using the training set,  $\mathcal{D}_t$ , as in (2). This training objective is known as empirical risk minimization (ERM). In other words, the ERM aims to minimize the average per instance training loss.

$$\text{ERM} = \min_{\theta} \frac{1}{N} \sum_{i=0}^N \ell(y_i, \{f_\theta(x_i)\}_{c=1}^C) \quad (2)$$

A popular  $\ell$  used in deep learning is the cross-entropy loss:

$$\ell(y_i, \{f_\theta(x_i)\}_{c=1}^C) = -\log p(y_i | x_i; \theta)$$

$$= -\log \left( \frac{\exp(f_\theta(x_i))}{\sum_{c=1}^C \exp(f_\theta(x_c))} \right) \quad (3)$$

The pseudocode for ERM is shown in Algorithm 1 below. In a classification problem, each output label  $y_i$  belongs to one of  $C$  classes and the aim of  $f_\theta(\cdot)$  is to ensure that each input  $x_i$  is classified to the correct class  $c \in \{1, \dots, C\}$ .

---

### Algorithm 1: Empirical Risk Minimization Training

---

**Input:** Training Data  $\mathcal{D}_t = \{(x_i, y_i)\}_{i=1}^N$ ; no of Epochs,  $E$ ; Batch size,  $b$ ; Learning rate  $\eta$   
**Output:** Trained Model  
 Initialise model parameters  $\theta_0$   
**for**  $t = 1, 2, \dots, E$  **do**  
     randomly split  $\mathcal{D}_t$  into  $n$  equal-sized minibatches;  $|B| = b$   
     **for**  $B \in \{1, \dots, n\}$  **do**  
         perform forward pass for model  $f_\theta$   
          $\nabla_{\theta} L(\theta_t) \leftarrow \frac{1}{|B|} \sum_{i=0}^b \nabla_{\theta} L(y_i, f_\theta(x_i))$  #Compute loss and  
         perform gradient step  
          $\theta_{t+1} \leftarrow w_t - \eta \cdot \nabla_{\theta} L(w_t)$  #Update model parameters in  
         backward pass  
     **end**  
**end**

---

Unlike ERM, where the average per sample loss over the entire training data is minimized, gDRO minimizes the worst group error. Thus, gDRO presumes group annotations over the training data i.e., every training sample is a triplet  $\{(x_1, y_i, g_1), \dots, (x_n, y_n, g_n)\}$  where the  $g_n$  stands for the group annotation of the  $n$ th sample. On the contrary, no group annotations based on spurious correlations are assumed in this study. Rather, a typical case of class imbalance where samples belonging to the minority class are fewer than those of the majority class is the focus of this study. Thus, number of samples belonging to each class is denoted as  $N_c$ . Hence, instead of (2), (4) is used to update the DNN in gDRO in this study; where  $\frac{1}{\sqrt{N_c}}$  is for the group adjustment as in [32].

$$\text{gDRO} = \min_{\theta} \max_{c \in C} \left\{ \frac{1}{N_c} \sum_{i=1}^{N_c} \ell(y_i, f_\theta(x_i)) + \frac{1}{\sqrt{N_c}} \right\} \quad (4)$$

The pseudocode for group DRO training is presented in Algorithm 2.

## III. MATERIALS AND METHODS

### A. Benchmark Datasets

Nineteen (19) carefully selected binary class benchmark imbalanced datasets from Keel<sup>1</sup> and UCI<sup>2</sup> data repositories are used in this study. Details such as the sample size, number of features, fraction of majority and minority samples in percentage are presented in Table I. The degree of imbalance in each dataset is indicated by the imbalance ratio (IR) which is the ratio of the majority to minority samples size. Likewise, the degree of complexity of each dataset is shown using mean silhouette coefficient of its samples (S.Coeff) [36]. The S.Coeff values ranges between -1 and +1. Values around zero indicate overlapping class clusters, whereas values close to +1 and -1

<sup>1</sup> <https://sci2s.ugr.es/keel/index.php>

<sup>2</sup> <https://archive.ics.uci.edu/ml/datasets.php>

indicate well separated and highly overlapped class clusters respectively.

**Algorithm 2:** Group Distributionally Robust Optimization Training

**Input:** Training Data  $\mathcal{D}_t = \{(x_i, y_i)\}_{i=1}^N$ ; classes,  $c \in \{1, \dots, C\}$ ; no of Epochs,  $E$ ; Batch size,  $b$

**Output:** Trained Model

Initialize parameters of model  $f_\theta$

for  $t = 1, 2, \dots, E$  do

    randomly split  $\mathcal{D}_t$  into  $n$  equal-sized minibatches;  $|B| = b$

    for  $B \in \{1, \dots, n\}$  do

        perform forward pass for model  $f_\theta$

        for  $c = \{1, \dots, C\}$  do

$L_c(\theta_t) \leftarrow \frac{1}{N_c} \sum_{i=1}^{N_c} \ell(y_i, f_\theta(x_i)) + \frac{1}{\sqrt{N_c}}$  # compute loss for each class

        end

$\nabla_\theta L(\theta_t) \leftarrow \nabla_\theta \{\max(L_c(\theta_t)_{c=1}^C)\}$  # perform gradient step with worst group

$\theta_{t+1} \leftarrow \theta_t - \eta \cdot \nabla_\theta L(\theta_t)$  # update model parameters in backward pass

    end

end

**B. Methods for Handling Class Imbalance**

In addition to ERM and gDRO, experiments were also carried out using four classical imbalance methods and compared. These methods were chosen to cover commonly used imbalanced methods in DNN research and their hybrid.

- Random Oversampling (ROS)
- Random Undersampling (RUS)
- Cost sensitive reweighting (COST): The weights assigned to majority and minority class samples are determined by the inverse of  $N_c$ ; where  $N_c$  is the number of samples belonging to class  $c$ .
- Hybrid of random undersampling and oversampling (RUSROS): This involves initially randomly undersampling the majority class by 50% before randomly oversampling the minority class samples till it equals the majority class size.

Overall, the performance of six methods (ERM, gDRO, ROS, RUS, COST and RUSROS) on the imbalance datasets are compared in this study.

**C. DNN Architecture**

Unlike convolutional neural networks (CNN), where a wide range of benchmark architectures are available [37], determining an appropriate DNN architecture for tabular data is nontrivial due to the sparsity of representative works addressing pertinent issues such as the ideal network depth and width as well as the best activation functions for this class of models. While in recent years several novel architectures have been proposed in representative studies [38-41], no single method provides a reliable performance across multiple tasks. Hence, deep fully connected otherwise known as deep multilayer perceptron remains the quintessential baseline architecture for modelling structured data [41] and thus, used in this study. Besides, deep fully connected neural networks are

natural fit for imbalanced data with capability of yielding impressive results when the hyperparameters are optimized [42].

TABLE I. BENCHMARK DATASETS

Data	# Samples	# Features	% Maj Class	% Min Class	IR	S.Coeff
abalone19	4174	8	99.23	0.77	129.44	-0.021
protein_homo	145751	74	99.11	0.89	111.46	0.556
mamography	11183	6	97.68	2.32	42.01	0.45
ozone_level	2536	72	97.12	2.88	33.74	-0.049
wine_quality	4898	11	96.26	3.74	25.77	0.146
oil	937	49	95.62	4.38	21.85	0.084
abalone	731	8	94.25	5.75	16.4	0.107
glass4	214	9	93.93	6.07	15.46	0.363
coverttype	38501	54	92.87	7.13	13.02	0.114
vowel0	988	13	90.89	9.11	9.98	0.166
satimage	6435	36	90.27	9.73	9.28	-0.134
page-blocks0	5472	10	89.78	10.2	8.79	0.505
ecoli3	336	7	89.58	10.4	8.6	0.126
segment0	2308	19	85.75	14.3	6.02	-0.063
yeast4	1484	8	83.56	16.4	5.08	0.037
vehicle0	846	18	76.48	23.5	3.25	0.065
haberman	306	3	73.53	26.5	2.78	0.069
phoneme	5404	5	70.65	29.4	2.41	0.087
pima	768	8	65.1	34.9	1.87	0.092

ReLU activation function is widely used in deep learning class imbalance research [14, 43], hence the same is adopted in the DNN model used in this study. Batch normalization and 0.5 dropout rate are applied after each ReLU activation of each hidden layer to avoid overfitting. Likewise, to optimize a DNN model for each dataset, representative imbalance studies have often resulted to grid search for hyperparameters optimization [14, 44, 45]. Similarly, 80% of each imbalanced data was used for hyperparameter optimization via grid search. Only the depth and width of each DNN model are optimized as in [14]. A network width of 50 neurons per hidden layer was found to be sufficient for the models to overfit the data after experimenting with widths of 512, 300, 100, 50 and 32 neurons respectively. The depth of these models was optimized starting with a depth of two (i.e., 2-hidden layers) and varying it up to six. The optimal DNN architecture for each imbalanced dataset

was determined via the best mean AUC over 5-fold cross validation [42]. The AUC results for optimal number of layers and architecture for each dataset is presented in Table IV of the appendix.

#### D. Experimental Setup

10-fold cross validation approach was employed for model training and validation for each combination of dataset, model and imbalance method. All DNN models were trained to stop early if there are no improvement in the validation error after 10 successive epochs. Adam stochastic optimization with learning rate of 0.001 was used in model training. Each model weights were randomly initialized with uniform distribution and Xavier variance [46] with zero bias before training with batch stochastic gradient descent. The minibatch sizes were set to range from 1/32 to 1/100 of each respective dataset sample sizes. The training procedure is illustrated in Fig. 1.

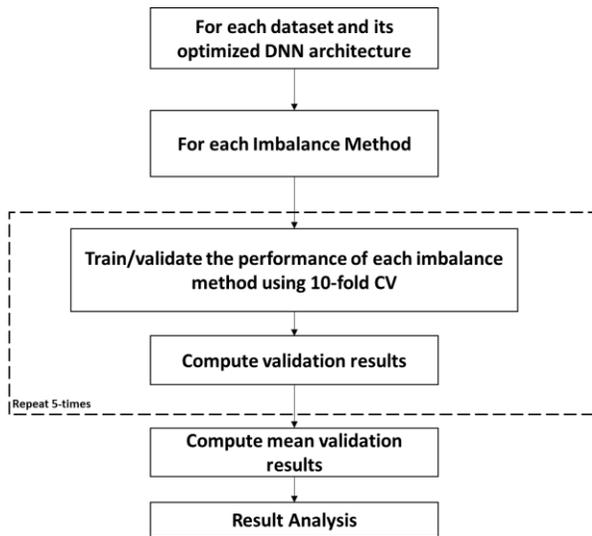


Fig. 1. Training procedure

To further ensure credibility of our findings and handle associated inconsistencies resulting from randomization and stochastic nature of the training process, the experiment for each combination was repeated five times with varying random seeds; resulting in different initial parameters for each repetition [47]. The mean validation results across the repetitions are reported. In all, 5,700 experiments were conducted. Given the large number of models that is required to be trained, a commodity hardware GPU, Nvidia Geforce GTX 960M, on a core i5 Dell Inspiron 7559 machine was leveraged to speed up the experiments. All experiments were implemented in Pytorch deep learning framework [48].

#### E. Evaluation Metrics

The inadequacy of accuracy and error rate as measures of classification performance of imbalance datasets is well documented in the literature [49]. Thus, four complementary evaluation metrics that have been used in imbalance learning research have been adopted in the study. Each of this metrics is described in what follows. Note that FP, FN, TP and TN are false positive, false negative, true positive and true negative respectively.

- Receiver Operating Characteristic Area Under the Curve (ROC-AUC): The ROC is a plot of the true positive rate ( $\frac{TP}{TP+FP}$ ) against the false positive rate ( $\frac{FP}{FP+FN}$ ) across all possible discrimination thresholds. From this plot, the AUC which is the area under the receiver operating characteristic (ROC) curve can be calculated and used as a performance measure of classification model.
- Precision-Recall AUC curve (PR-AUC), perhaps inspired by the ROC-AUC, is a plot of precision on the y-axis ( $\frac{TP}{TP+FP}$ ) against recall ( $\frac{TP}{TP+FN}$ ) on the x-axis. The area under the PR-curve is also used as a measure of the performance of binary classification models. The AUC implementation used in this work is calculated using the trapezoidal rule.
- F1-Measure, a widely used evaluation metric, is another measure of evaluation used in this study. It is the harmonic mean of precision and recall (i.e.,  $\frac{(1+\beta^2) \times Recall \times Precision}{\beta^2 \times Recall + Precision}$ ). The beta ( $\beta$ ) parameter shows the trade-off between precision and recall. Our interest is to detect both majority and minority classes with equal preference, hence, the  $\beta$  parameter is set to 1.
- Geometric mean (g-mean) is the final evaluation metric used to evaluate the models. In case of binary classification, g-mean is the squared-root of the product of recall and true negative rate (TNR) ( $\sqrt{\frac{TP}{TP+FN} \times \frac{TN}{TN+FP}}$ ).

Additionally, the Friedman test [50] is also used in this study to detect differences in the experimental results across multiple attempts, when the normality assumption may not hold. Thus, this test is used to reject the null hypothesis that the compared methods produce similar performance across the different datasets and DNNs in comparison to their mean rankings. Then, as recommended in [51], pairwise posthoc analysis using Wilcoxon signed-rank test [52] with Holm's alpha (0.05) correction [53] was used for comparison. A visualization of the comparison is presented using a critical difference diagram [54].

## IV. RESULTS

Multiple benchmark datasets enabled a fair comparative analysis of ERM, gDRO and the imbalance methods across a wide range of imbalance ratios. The mean ( $\pm$ standard deviation) of the validation performance of each method across the respective evaluation metrics for each dataset over five repetitions of the experiment (as shown in Fig. 1) are presented in Fig 2. The overall average performance of each method in addition to the number of times each method ranks first for each evaluation metric is also presented in Table II. Similarly, a bar plot showing the average ranking of each method per dataset is presented in Fig. 3.

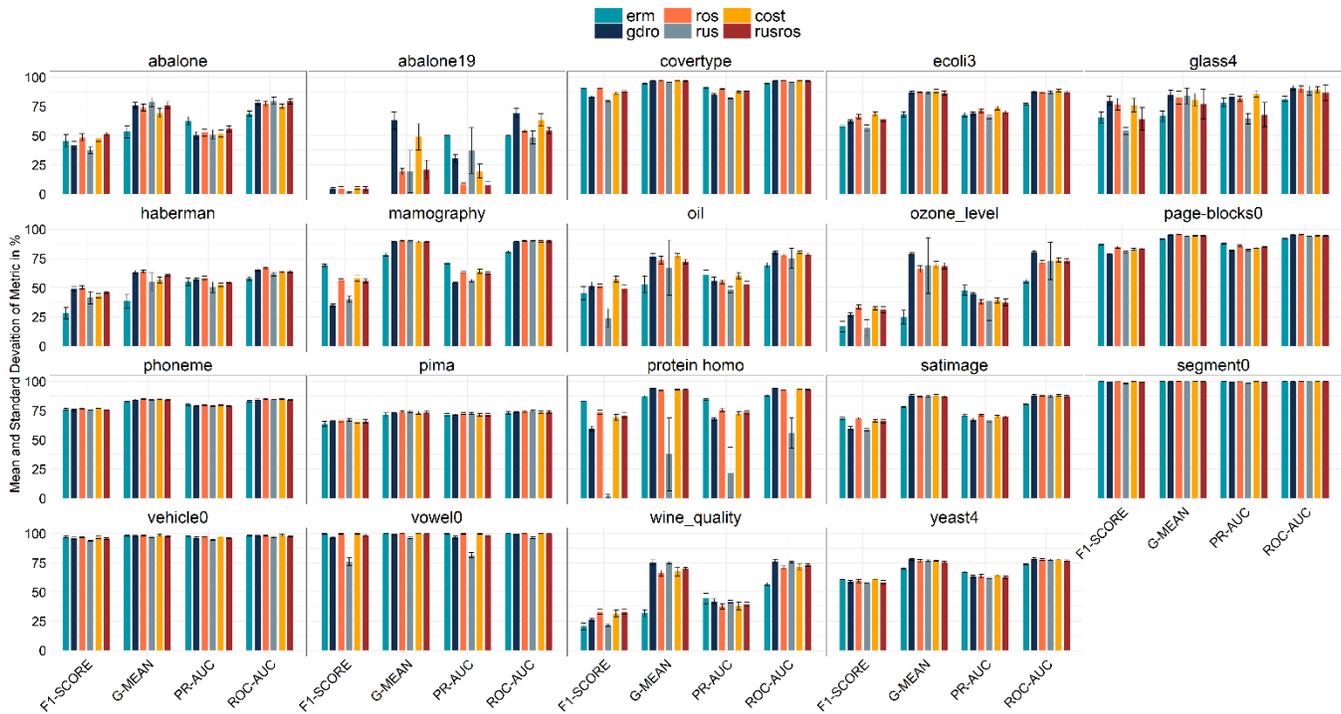


Fig. 2. Mean ( $\pm$ standard deviation) results of comparative methods

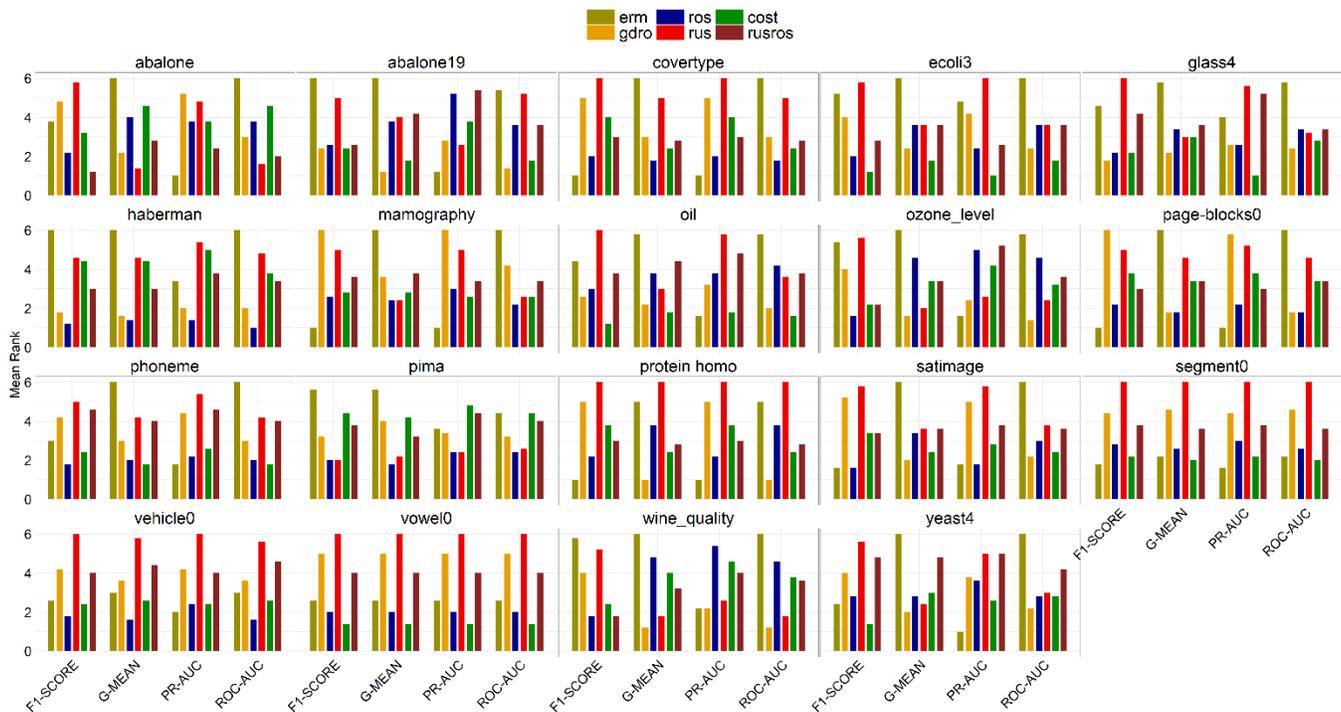


Fig. 3. Mean ranking of the experimental results (lower the better)

### A. Results based on F1-Score

As, depicted in Fig. 2, the performance of gDRO in comparison to ERM regarding the top six imbalanced benchmarks (IR of 129.44 to 21.85) shows that gDRO mostly produces better average f1-score for those with overlapping classes (abalone19, ozone level, wine\_quality and oil) i.e.,

gDRO identifies minority samples better on complex imbalanced datasets- whereas ERM performed better on those with better class separability (protein\_homo and mamography). The general f1-score on the abalone19 benchmark is notably very poor across the compared methods, a likely explanation for this is lack of data/information (given a meagre minority sample size of 32) worsened by class overlap.

Likewise, the superior performance of gDRO over ERM on this benchmark suggests the suitability of the former in extreme imbalance cases where data is lacking. As for the remaining datasets in decreasing order of imbalance ratio, ERM produced better average f1-score than gDRO in all but glass4, ecoli3 and haberman datasets; the three smallest datasets in the benchmarks and of varying degree of overlap. In comparison to the classical imbalance methods, while mostly outperforming RUS on mammography, page-block0 and pima, gDRO mostly produced lower average f1-score than, at least, one of ROS, COST or RUSROS except on abalone19 and glass4 datasets where gDRO outperformed other methods. As shown in Table II, ROS produced the best average overall f1-score across the benchmarks, producing the best score seven times in the process. On the other hand, RUS produced the lowest overall f1-score and only managed to produce the best score once (which was jointly with ROS on the least imbalance dataset, pima).

### B. Results based on G-Mean

As shown in Fig. 2, gDRO produced better average g-mean scores than ERM across all but the segment0, vehicle0 and vowel0 datasets where they both performed similarly or marginally better performance by ERM. It should be noted that the performance of the model on these datasets is generally better than the remaining datasets. In comparison to the selected imbalance methods in terms of the top six imbalanced datasets (abalone19, protein homo, mammography ozone level, wine\_quality and oil), gDRO yielded the best average g-mean on all but the mammography and oil datasets where ROS and RUS as well as COST respectively produced better performance. On the remaining datasets, as the imbalance ratio reduces and the minority to majority ratio increases, the performance of the classical imbalance methods generally become more comparable to gDRO which was only able to produce the best average g-mean on the glass4, satimage, page-block0 and yeast4 datasets. Table II shows that gDRO produced the best average overall g-mean of 84.28% across the benchmarks, while achieving the best score 8 times in the process. The implication of the impressive performance of gDRO in terms of g-mean is that it detects the minority samples with lower false positive and false negative rates. On the other hand, an overall average g-mean score of 67.74% achieved by ERM makes it the least performing method in this regard.

### C. Results based on PR-AUC

As shown in Fig. 2, ERM outperforms gDRO and other classical imbalance methods based on average pr-auc on all but the ecoli3, glass4, haberman and pima datasets. Three of these datasets (ecoli3, glass4 and haberman) have the least number of samples while the relatively larger pima dataset is the least imbalance amongst the considered benchmarks. Despite achieving better average pr-auc than ERM on these datasets, the performance of gDRO still remains inferior to at least one of ROS, RUS, COST or RUSROS. Table II further shows that ERM produced the best overall average pr-auc of 73.07%, achieving the best performance 14 times across the different benchmarks. In contrast, RUS showed the lowest performance in this regard despite producing the best result once (jointly with ROS on the pima dataset).

TABLE II. AVERAGE PERFORMANCE ACROSS ALL DATASETS (AND THE NUMBER OF TIMES EACH METHOD RANKS FIRST)

Metric	ERM	GDRO	ROS	RUS	COST	ROSRUS
F1-SCORE	61.85 (6)	60.57 (2)	65.07 (7)	51.54 (1)	64.37 (5)	63.19 (2)
G-MEAN	67.74 (0)	84.28 (8)	80.36 (6)	77.23 (2)	81.68 (5)	80.12 (0)
PR-AUC	73.07 (14)	68.16 (1)	68.41 (3)	62.83 (1)	68.95 (3)	66.92 (0)
ROC-AUC	77.86 (0)	85.55 (8)	83.80 (4)	81.10 (0)	84.28 (0)	83.58 (1)

### D. Results based on ROC-AUC

The performance of gDRO as illustrated in Fig. 2 relative to ERM in terms of average roc-auc is similar to findings based on average g-mean as gDRO shows better average roc-auc scores than ERM across all but the segment0, vehicle0 and vowel0 datasets where they both performed similarly or ERM is marginally better. Comparison based on the top six imbalanced benchmarks also shows similar trend as gDRO produced the best average roc-auc on all but the mammography (where ROS, RUS, COST and RUSROS were better) and oil (where COST was better) datasets. Datasets with lower IR produced improved results that are comparable or relatively better than gDRO with these methods. However, gDRO produced the best average roc-auc on the glass4, satimage, page-block0 and yeast4 datasets. Further, Table II shows that gDRO produced the highest overall average roc-auc (85.55%), producing the results eight times across the datasets. On the other hand, ERM produced the lowest overall average performance based on ROC-AUC.

### E. Statistical Analysis

The Friedman's test results presented in Table III shows that the null hypothesis, namely, the imbalance methods produce similar performance across the different datasets is rejected. Hence, pairwise posthoc analysis using Wilcoxon signed-rank test is carried out to rank each method across the evaluation metrics as illustrated in Fig. 4. The figure shows a critical difference diagram of the imbalance methods for each evaluation metric where a thick horizontal line indicates group of imbalance methods (a clique) for which the difference in their performance is not statistically significant. For each metric, the mean performance of each method across the dataset is used for the statistical test. The figure shows that although for f1-score, g-mean, pr-auc and roc-auc, ROS, gDRO, ERM and gDRO respectively rank highest, each of these methods is not significantly better than at least two other methods. For instance, in terms of f1-score, ROS is not statistically different from COST and ERM whereas in term of pr-auc, no statistically significant difference exists between ERM, ROS, COST and gDRO.

TABLE III. RESULT OF FRIEDMAN'S TEST

Metric	p-value
F1-Score	0.00000
G-Mean	0.00000
PR-AUC	0.00000
ROC-AUC	0.00000

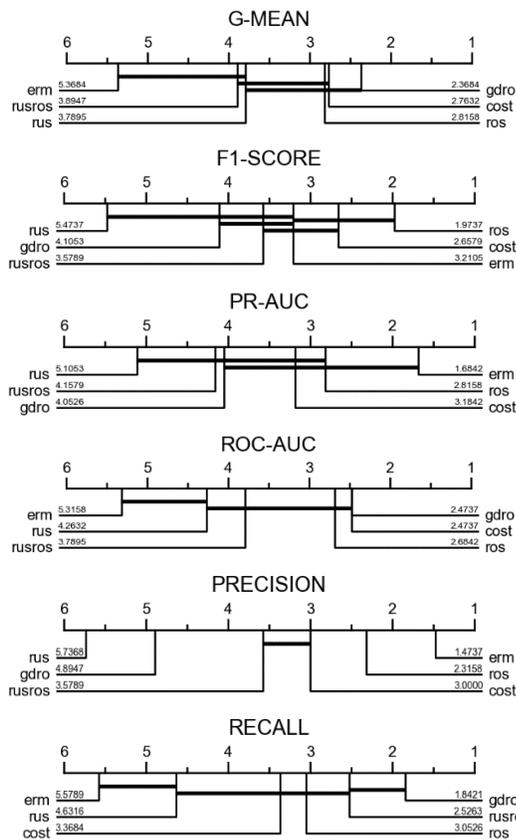


Fig. 4. Critical difference diagram of pairwise statistical difference comparison between imbalance methods

## V. DISCUSSION

The importance of this study cannot be overemphasized as it investigates gDRO for learning imbalance tabular data in DNN as against ERM which is the main learning objective in both balanced and imbalanced deep learning research. The empirical results of this study show that in terms of ROC-AUC which reflects a balanced assessment of the performance of the DNN models on both majority and minority samples across different thresholds, gDRO mostly outperforms ERM and most of the compared imbalance methods on benchmark datasets of varying imbalance ratios, sizes and complexities. This implies that deep imbalance learning via minimization of worst-case loss across classes can produce models that are more robust to both majority and minority class samples than ERM. Hence, gDRO is an ideal training objective in cases where both majority and minority classes of the imbalance data are equally important as it is known for its robustness to distributional shifts [32].

Following the generally held notion that ERM biases DNN models to the majority class when learning imbalance data [55], it would be expected that ERM perform poorly on the minority samples. However, it can be observed that ERM mostly outperform gDRO and other compared imbalance methods across the studied benchmarks in terms of pr-auc which mainly focuses on the performance of the DNN model on the minority samples under a range of thresholds. One likely explanation for this is that while ERM produced the best

overall result on the minority samples over different thresholds, this does not necessarily mean it produced the best performance for a specific threshold value [56]. This explains why ERM is unable to replicate similar feat on metrics like f1-score and g-mean that are computed for specific threshold value.

Additionally, since pr-auc measures the area under the plot of precision against recall for different thresholds, another possible explanation could be that the pr-auc of ERM is dominated by precision. As hinted in Section E of IV, precision quantifies the number of correct positive (minority) predictions by dividing the number of correctly classified positive samples by the total number of correctly classified positive samples and negative (majority) samples that are incorrectly classified as positive. Compared to recall which quantifies the number of correct positive predictions from all positive samples, only a model that is bias to the majority samples is less likely to misclassify a majority sample as minority than misclassify a minority sample. In other words, ERM could produce the best pr-auc while being bias to the majority class samples if pr-auc is dominated by precision. The critical difference diagrams for precision and recall have been included in Fig. 4 to further elucidate this line of thought. See Fig. 5 and Fig. 6 of the appendix for more details of each method on each dataset in terms of precision and recall.

In relation to the classical imbalance methods, COST and ROS tend to perform similarly and better than RUS in most cases with COST showing superior performance in highly imbalance cases. However, the performance gains in terms of minority sample detection for these methods tend to come at the expense of some majority class samples [20]. On the other hand, RUSROS does not appear to have any major advantage that its constituent methods have not exhibited. Generally, the inferior performance of the imbalance methods on the highly imbalance benchmarks with some degree of class overlap like abalone19 and ozone\_level compared to similarly imbalance ones like protein\_homo and mamography with more minority samples and better class separability underlines the impact of size and complexity of data in imbalance learning. Decreasing imbalance ratio tend to result in improved performance across the metrics.

It should however be noted that the adopted experimental design could have impacted the empirical results of this study. Particularly, in relation to reporting the mean scores of several repetitions of the experiments as against a single round as common in most DNN-based imbalance research [5]. Nevertheless, the adopted design has obvious benefits amongst which is an objective measure of the true model performance.

In sum, the choice of method for handling class imbalance in DNN models depends on several factors, including the severity of class imbalance, the size and complexity of the dataset, and the specific evaluation metric of interest. In some cases, ERM may be sufficient to achieve good performance on imbalanced datasets, while in other cases, gDRO or other methods may be necessary.

## VI. CONCLUSION

Deep imbalance learning has mainly focused on imbalance in computer vision related tasks. Likewise, empirical risk minimization (ERM) which entails minimizing the average per sample training loss over the entire training data has been shown in pertinent works to bias DNNs to the majority class when learning from imbalance data [31]. An alternative learning objective, group distributionally robust optimization (gDRO) is investigated in this study for imbalance learning with a focus on tabular data. The performance of gDRO in comparison with ERM and four classical imbalance resolution methods on several benchmark imbalance datasets of varying imbalance ratios are examined using four common metrics for evaluating class imbalance. Experimental findings show that while gDRO outperform other methods in terms of g-mean and ROC-AUC, whereas ERM and ROS rank highest in terms of pr-auc and f1-score respectively. Future research efforts will focus on the impact of pretraining on deep imbalance learning as well as gDRO for multiclass imbalance tabular data.

## REFERENCES

- [1] Bengio, Y., *Practical recommendations for gradient-based training of deep architectures*, in *Neural networks: Tricks of the trade*. 2012, Springer. p. 437-478.
- [2] Goodfellow, I., Y. Bengio, and A. Courville, *Deep learning*. 2016: MIT press.
- [3] Buda, M., A. Maki, and M.A. Mazurowski, A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 2018. 106: p. 249-259.
- [4] Geng, Y. and X.Y. Luo, Cost-sensitive convolutional neural networks for imbalanced time series classification. *Intelligent Data Analysis*, 2019. 23(2): p. 357-370.
- [5] Khoshgoftaar, J.M.J.M., Survey on deep learning with class imbalance. *Journal of Big Data*, 2019. 6(1).
- [6] Zhang, C., et al., A Cost-Sensitive Deep Belief Network for Imbalanced Classification. *IEEE Transactions on Neural Networks and Learning Systems*, 2019. 30(1): p. 109-122.
- [7] Arizmendi, C., et al., Automated classification of brain tumours from short echo time in vivo MRS data using Gaussian Decomposition and Bayesian Neural Networks. *Expert Systems with Applications*, 2014. 41(11): p. 5296-5307.
- [8] Afzal, S., et al., A Data Augmentation-Based Framework to Handle Class Imbalance Problem for Alzheimer's Stage Detection. *IEEE Access*, 2019. 7: p. 115528-115539.
- [9] Moepya, S.O., S.S. Akhoury, and F.V. Nelwamondo. Applying Cost-Sensitive Classification for Financial Fraud Detection under High Class-Imbalance. in *2014 IEEE International Conference on Data Mining Workshop*. 2014.
- [10] Thennakoon, A., et al. Real-time Credit Card Fraud Detection Using Machine Learning. in *2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*. 2019.
- [11] Fiore, U., et al., Using generative adversarial networks for improving classification effectiveness in credit card fraud detection. *Information Sciences*, 2019. 479: p. 448-455.
- [12] Choi, S., et al., Cost-sensitive Prediction of Airline Delays Using Machine Learning, in *2017 IEEE/AIAA 36th Digital Avionics Systems Conference*. 2017.
- [13] Mustapha, I.B., S.M. Shamsuddin, and S. Hasan, A Preliminary Study on Learning Challenges in Machine Learning-based Flight Delay Prediction. *International Journal of Innovative Computing*, 2019. 9(1).
- [14] Johnson, J.M. and T.M. Khoshgoftaar. Deep Learning and Data Sampling with Imbalanced Big Data. in *2019 IEEE 20th International Conference on Information Reuse and Integration for Data Science (IRI)*. 2019.
- [15] Ali-Gombe, A. and E. Elyan, MFC-GAN: Class-imbalanced dataset classification using Multiple Fake Class Generative Adversarial Network. *Neurocomputing*, 2019. 361: p. 212-221.
- [16] Wang, C., et al. CGAN-plankton: Towards large-scale imbalanced class generation and fine-grained classification. in *2017 IEEE International Conference on Image Processing (ICIP)*. 2017.
- [17] Khan, S.H., et al., Cost-Sensitive Learning of Deep Feature Representations From Imbalanced Data. *IEEE Transactions on Neural Networks and Learning Systems*, 2018. 29(8): p. 3573-3587.
- [18] Li, F.M., et al., Cost-Effective Class-Imbalance Aware CNN for Vehicle Localization and Categorization in High Resolution Aerial Images. *Remote Sensing*, 2017. 9(5).
- [19] Li, J.J.L.Z.L.W.L.L.W.Z.C.L.L.L.C.C., Automatic diagnosis of imbalanced ophthalmic images using a cost-sensitive deep convolutional neural network. *BioMedical Engineering OnLine*, 2017. 16(1).
- [20] Cui, Y., et al. Class-balanced loss based on effective number of samples. in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019.
- [21] Raj, V., S. Magg, and S. Wermter. Towards effective classification of imbalanced data with convolutional neural networks. in *IAPR Workshop on Artificial Neural Networks in Pattern Recognition*. 2016. Springer.
- [22] Zhang, C., K.C. Tan, and R. Ren. Training cost-sensitive Deep Belief Networks on imbalance data problems. in *2016 International Joint Conference on Neural Networks (IJCNN)*. 2016.
- [23] Cao, K., et al., Learning Imbalanced Datasets with Label-Distribution-Aware Margin Loss. *arXiv preprint arXiv:1906.07413*, 2019.
- [24] Kukar, M. and I. Kononenko. Cost-sensitive learning with neural networks. in *ECAI*. 1998. Citeseer.
- [25] Byrd, J. and Z. Lipton. What is the effect of importance weighting in deep learning? in *International Conference on Machine Learning*. 2019. PMLR.
- [26] Huang, C., et al. Learning Deep Representation for Imbalanced Classification. in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016.
- [27] Jia, F., et al., Deep normalized convolutional neural network for imbalanced fault classification of machinery and its understanding via visualization. *Mechanical Systems and Signal Processing*, 2018. 110: p. 349-367.
- [28] Mikolov, T., et al. Distributed representations of words and phrases and their compositionality. in *Advances in neural information processing systems*. 2013.
- [29] Chou, H.-P., et al. Remix: rebalanced mixup. in *European Conference on Computer Vision*. 2020. Springer.
- [30] Ren, J., et al., Balanced meta-softmax for long-tailed visual recognition. *arXiv preprint arXiv:2007.10740*, 2020.
- [31] Zhang, H., et al., mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.
- [32] Sagawa, S., et al., Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019.
- [33] Ye, H.-J., et al., Identifying and compensating for feature deviation in imbalanced deep learning. *arXiv preprint arXiv:2001.01385*, 2020.
- [34] Duchi, J.C., T. Hashimoto, and H. Namkoong, Distributionally robust losses against mixture covariate shifts. Under review, 2019. 2.
- [35] Oren, Y., et al., Distributionally robust language modeling. *arXiv preprint arXiv:1909.02060*, 2019.
- [36] Rousseeuw, P.J., Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 1987. 20: p. 53-65.
- [37] Ding, W., et al. Facial action recognition using very deep networks for highly imbalanced class distribution. in *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. 2017.
- [38] Borisov, V., et al., Deep neural networks and tabular data: A survey. *arXiv preprint arXiv:2110.01889*, 2021.
- [39] Arik, S.O. and T. Pfister, Tabnet: Attentive interpretable tabular learning. *arXiv preprint arXiv:1908.07442*, 2019.

[40] Huang, X., et al., Tabtransformer: Tabular data modeling using contextual embeddings. arXiv preprint arXiv:2012.06678, 2020.

[41] Gorishniy, Y., et al., Revisiting Deep Learning Models for Tabular Data. arXiv preprint arXiv:2106.11959, 2021.

[42] Díaz-Vico, D., A.R. Figueiras-Vidal, and J.R. Dorronsoro. Deep MLPs for Imbalanced Classification. in 2018 International Joint Conference on Neural Networks (IJCNN). 2018.

[43] Glorot, X., A. Bordes, and Y. Bengio. Deep sparse rectifier neural networks. in Proceedings of the fourteenth international conference on artificial intelligence and statistics. 2011. JMLR Workshop and Conference Proceedings.

[44] Wang, S.J., et al. Training Deep Neural Networks on Imbalanced Data Sets. in 2016 International Joint Conference on Neural Networks. 2016.

[45] Hu, Z.X. and P. Jiang, An Imbalance Modified Deep Neural Network With Dynamical Incremental Learning for Chemical Fault Diagnosis. Ieee Transactions on Industrial Electronics, 2019. 66(1): p. 540-550.

[46] Glorot, X. and Y. Bengio. Understanding the difficulty of training deep feedforward neural networks. in Proceedings of the thirteenth international conference on artificial intelligence and statistics. 2010. JMLR Workshop and Conference Proceedings.

[47] Pant, H., M. Sharma, and S. Soman, Twin neural networks for the classification of large unbalanced datasets. Neurocomputing, 2019. 343: p. 34-49.

[48] Paszke, A., et al., Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems, 2019. 32: p. 8026-8037.

[49] Branco, P., L.s. Torgo, and R.P. Ribeiro, A Survey of Predictive Modeling on Imbalanced Domains. ACM Comput. Surv., 2016. 49(2): p. Article 31.

[50] Friedman, M., A comparison of alternative tests of significance for the problem of m rankings. The Annals of Mathematical Statistics, 1940. 11(1): p. 86-92.

[51] Benavoli, A., G. Corani, and F. Mangili, Should we really use post-hoc tests based on mean-ranks? The Journal of Machine Learning Research, 2016. 17(1): p. 152-161.

[52] Wilcoxon, F., Individual comparisons by ranking methods. 1992: Springer.

[53] Holm, S., A simple sequentially rejective multiple test procedure. Scandinavian journal of statistics, 1979: p. 65-70.

[54] Demšar, J., Statistical comparisons of classifiers over multiple data sets. The Journal of Machine Learning Research, 2006. 7: p. 1-30.

[55] Ye, H.-J., D.-C. Zhan, and W.-L. Chao. Procrustean training for imbalanced deep learning. in Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021.

[56] Davis, J. and M. Goadrich. *The relationship between Precision-Recall and ROC curves.* in *Proceedings of the 23rd international conference on Machine learning.* 2006.

APPENDIX

TABLE IV. ROC-AUC SCORES FOR OPTIMAL ARCHITECTURE SELECTION FOR EACH DATASET

Dataset	2_lyrs	3_lyrs	4_lyrs	5_lyrs	6_lyrs
abalone19	50	50	50	50	<b>50</b>
abalone	66.33358	64.81243	71.23284	<b>71.59815</b>	66.15334
covertype	93.75338	<b>94.1079</b>	93.28036	93.94478	93.53545
ecoli3	71.14201	64.17092	69.55867	71.43367	<b>75.80867</b>
glass4	79.375	74.6875	74.6875	74.6875	<b>84.6875</b>
haberman	<b>65.53534</b>	61.76715	62.7362	63.29175	61.96581
mamography	74.85703	76.7389	77.4536	<b>78.67885</b>	78.64452
oil	69.72028	72.29672	<b>72.50651</b>	71.21878	61.21878
ozone_level	<b>54.46271</b>	50.76923	50	51.64122	52.2568
page-blocks0	90.47406	90.94058	90.8244	<b>91.70862</b>	90.76801
phoneme	82.28141	83.61735	83.42348	83.15185	<b>85.3148</b>
pima	72.05325	72.35902	73.29717	72.05595	<b>73.51786</b>
protein_homo	85.52046	86.38106	<b>88.44263</b>	87.1977	87.19304
satimage	78.53431	<b>80.78899</b>	78.5861	79.28501	78.04006
segment0	<b>99.96845</b>	99.77978	99.74823	99.5911	99.74823
vehicle0	96.21446	95.72569	96.74197	95.51737	<b>96.86146</b>
vowel0	99.93056	99.93056	99.21627	99.14683	<b>100</b>
wine_quality	54.41053	<b>56.35424</b>	54.12273	52.58129	52.51172
yeast4	73.4827	73.3687	73.96551	72.27256	<b>75.26621</b>

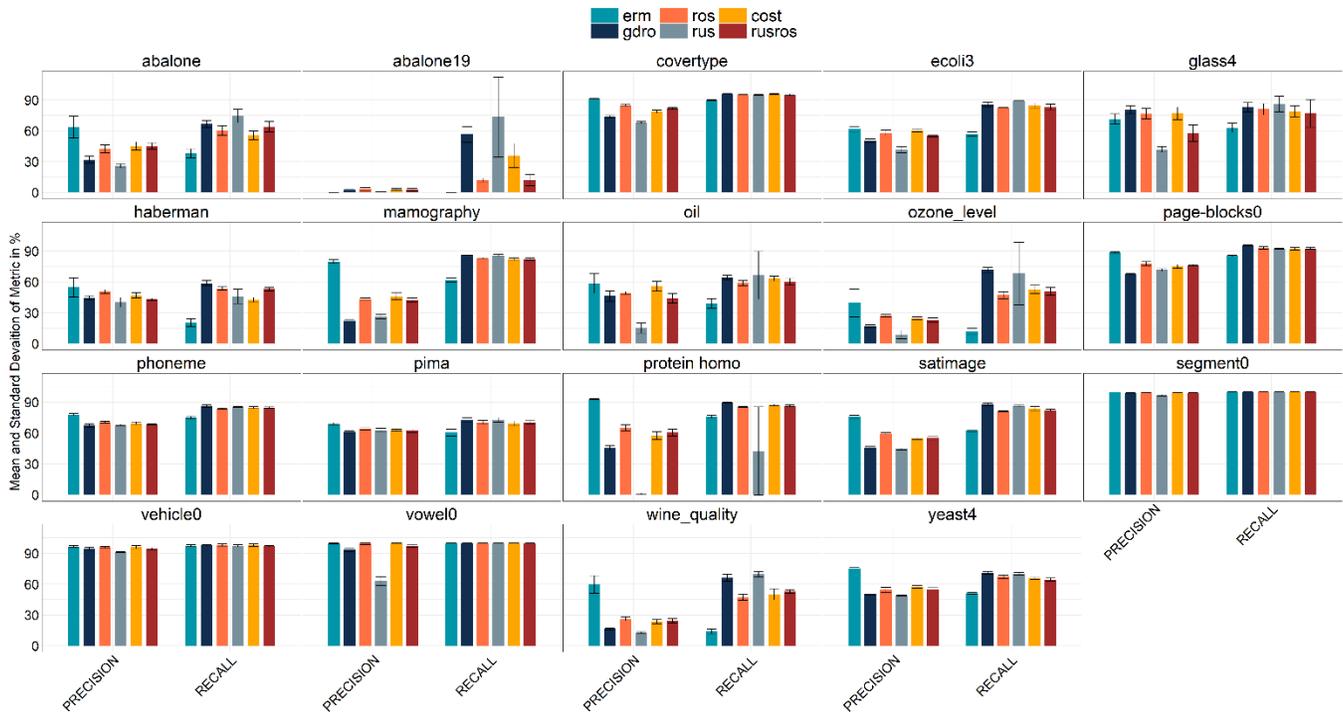


Fig. 5. Mean ( $\pm$ standard deviation) for precision and recall of comparative methods

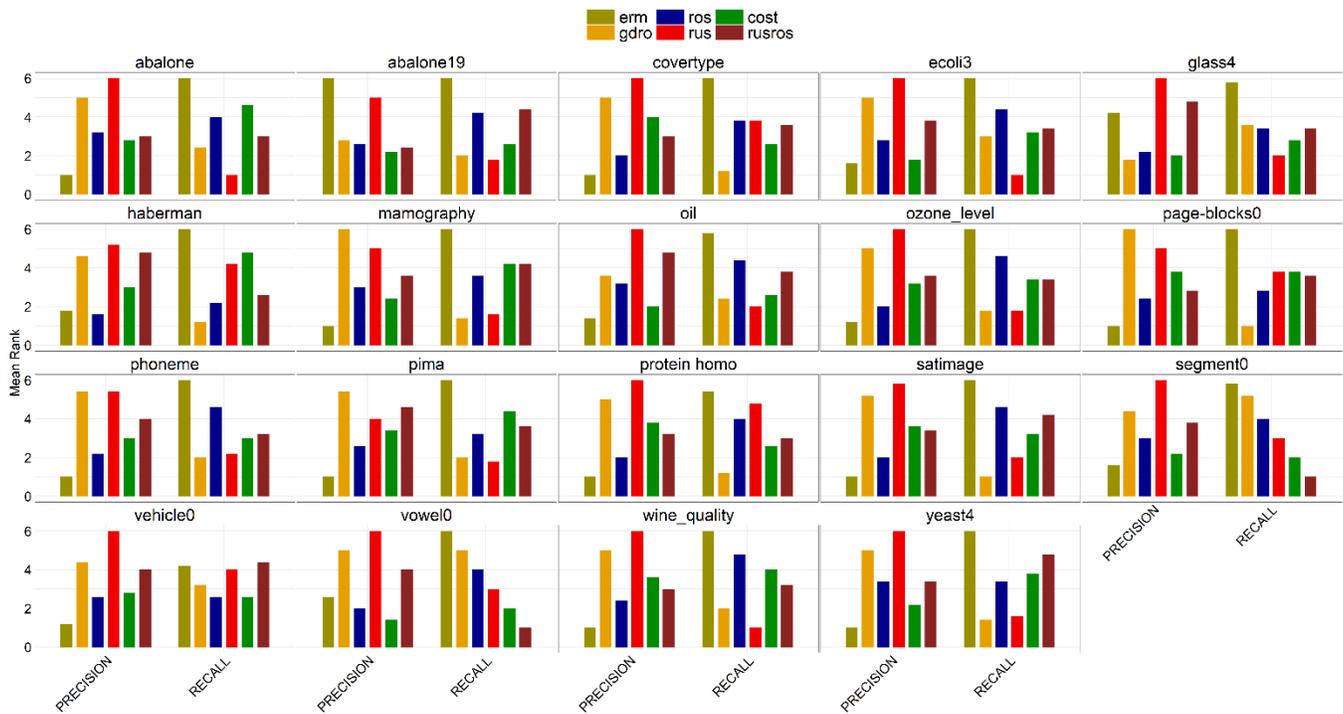


Fig. 6. Mean ranking precision and recall (lower the better)

# A Biologically Inspired Appearance Modeling and Sample Feature-based Approach for Visual Target Tracking in Aerial Images

Lili Pei<sup>\*1</sup>, Xiaohui Zhang<sup>2</sup>

College of Architectural Engineering, Tangshan Polytechnic College, Tangshan, 063299, China<sup>1</sup>  
College of Information Engineering, Tangshan Polytechnic College, Tangshan, 063299, China<sup>2</sup>

**Abstract**—Visual tracking in uncrewed aerial vehicles is challenging because of the target appearance. Various research has been fulfilled to overcome appearance variations and unpredictable moving target issues. Visual saliency-based approaches have been widely studied in biologically inspired algorithms to detect moving targets based on attentional regions (ARs) extraction. This paper proposes a novel visual tracking method to deal with these issues. It consists of two main phases: spatiotemporal saliency-based appearance modeling (SSAM) and sample feature-based target detection (SFTD). The proposed method is based on a tracking-by-detection approach to provide a robust visual tracking system under appearance variation and unpredictable moving target conditions. Correspondingly, a semi-automatic trigger-based algorithm is proposed to handle the phases' operation, and a discriminative-based method is utilized for appearance modeling. In the SSAM phase, temporal saliency extracts the ARs and coarse segmentation. Spatial saliency is utilized for the object's appearance modeling and spatial saliency detection. Because the spatial saliency detection process is time-consuming for multiple target tracking conditions, an automatic algorithm is proposed to detect the region saliences in a multithreading implementation that leads to low processing time. Consequently, the temporal and spatial saliencies are integrated to generate the final saliency and sample features. The generated sample features are transferred to the sample feature-based target detection (SFTD) phase to detect the target in different images based on samples. Experimental results demonstrate that the proposed method is effective and presents promising results compared to other existing methods.

**Keywords**—Visual tracking; biologically inspired; visual saliency detection; appearance modeling; attention region; spatiotemporal

## I. INTRODUCTION

Visual tracking is mainly investigated as an active research topic according to its wide range of applications, including smart surveillance systems, intelligent remote sensing technologies, action recognition, and robotic and human-computer interaction [1]. Particularly, visual target tracking is broadly studied on uncrewed aerial vehicles (UAV) to detect and track targets on aerial images. As opposed to applications with fixed cameras, for example, traffic monitoring, aerial videos have the favorable circumstances of higher portability and superior reconnaissance and surveillance [2]. However, visual tracking systems are still suffered from various challenges and difficulties. Target appearance variations and

tracking the target under uncontrollable and unpredictable conditions are as main challenges of these systems [3, 4]. In this regard, online learning-based tracking methods that can incrementally update feature representations have received more attention for achieving a reliable and robust visual tracking algorithm [5]. Therefore, an online target feature representation is crucial for preserving an efficient appearance model to describe and identify the target from background [3, 6].

Recently, biologically inspired and cognition-based approaches have become a topic of interest by many researchers [7]. These approaches are inspired by human biological mechanisms, meaningfully indicating that human perception is sensitive to attentional regions (ARs) [5, 8]. By adopting biologically inspired approaches, visual saliency detection methods have been presented to detect attentional regions based on spatial and temporal information, resulting in the design of a significant number of saliency-based object detection methods [2]. Hence, a fascinating question is how we can exploit and take advantage of these approaches to develop a more powerful visual tracking algorithm [9]. Current visual saliency detection approaches can be categorized as task-driven attention (top-down) and data-driven attention (bottom-up) [10-13]. The top-down approach is a result of long-term visual simulation with prior knowledge [10]. The top-down approaches focused on high-level information investigation, such as the sky, faces, and humans [14]. This approach's drawback is a hard generalization because it is not simply obtained from images [15].

Furthermore, they are slow and computationally expensive [10]. On the other hand, the bottom-up approaches are based on low-level visual features simulating the formation of short-term visual attention [16]. In contrast to the top-down method, the bottom-up-approaches are rapid [14, 17]. As promising results indicated on bottom-up based approaches, this study focuses on the bottom-up approach.

## II. RELATED WORKS

In this section, existing related works discuss two categories, visual saliency-based, and appearance modeling-based methods.

### A. Visual Saliency-Based Approach

As discussed previously, saliency-based methods are categorized as bottom-up and top-down methods. The bottom-

up methods are categorized into temporal, spatial, and combined-based approaches [2]. The details of each approach discuss to address the advantages and disadvantages.

1) *Temporal saliency*: Detection of salient regions is highly dependent on the recognition of moving objects since motions attract more attention [18]. For moving object detection from a video, moving object detection methods are mainly based on temporal information, such as background subtraction [19, 20], frame difference [21, 22], and optical flow [23].

Background subtraction is based on background modeling. It is widely used for moving object detection. It segments foreground objects from the image background to detect objects that are not moving [24]. However, the background subtraction method suffers from some limitations. These methods are sensitive to the fixed background, and the object extraction fails when the background has changed [10].

Optical flow is also used for moving object detection, even under moving camera conditions. However, optical flow is computationally expensive and sensitive to noise. Thus, optical flow algorithms are not robust in real-time visual tracking systems [2, 25].

Frame differencing as a practical approach for moving object detection is based on pixel-wise difference extraction among the image frames. The frame difference does not require background modeling and is not sensitive to fixed background-like background subtraction methods. Frame differencing is adaptive to dynamic backgrounds and has a low computational cost [2]. However, one of the major drawbacks of frame differencing is moving the objects during the frame capturing process. Because a target may have unpredictable motions, such as stop-and-go periods, the frame difference method is not robust under uncontrollable and unpredictable target movement conditions [26]. Therefore, it is required to propose a temporal saliency detection method to overcome the mentioned challenges [25].

2) *Spatial saliency*: Spatial saliency detection is based on low-level feature representation and focuses on salient region extraction from images. These features are investigated to describe and identify the region of interest as salient regions [2,40]. Several spatial-based methods have been developed, such as region low rank matrix recovery [27], region covariance [28, 29], color-based [41], contrast-based [30], frequency domain [31] and graph [32] methods. As mentioned earlier, visual object tracking in aerial has difficulties in target appearance variations and background changes. To deal with difficulties, spatial saliency detection can extract the salient regions that are moving targets. This method is not sensitive to background changes and abrupt motion.

3) *Spatiotemporal saliency*: Spatiotemporal saliency detection methods calculate the temporal and spatial saliencies [10] separately. They used motion cues investigation for moving object detection. Generally, the results only based on motion cues are not undesirable because of the lack of spatial distribution [2]. Therefore, it is required to integrate the

temporal and spatial features to detect the salient regions more accurately [9, 10,46].

### B. Appearance Modeling-based Approach

Appearance modeling approaches are normally used to deal with appearance variations challenges. These approaches are categorized into generative and discriminative-based methods [5].

1) *Generative-based methods*: Generative-based methods are used to generate a model of an object during appearance changes in scenes. The generated model exploits the discriminative features to handle the target's appearance variations. The mechanism of appearance model generation is frequently updated online to describe the appearance variations. Some generative-based methods are as follows. Lee and Kriegman [33] proposed a generative-based algorithm to update a model for target detection dynamically. Wu and Wang [34] proposed a real-time generative method integrated with an incrementally updating covariance modeling approach for visual tracking. Tianxiang and Li [35] presented an appearance modeling approach based on a generative method and structured sparse representation for tracking an object in a video. However, even though various generative-based methods have been proposed, these methods still have not fully exploited spatial identification within the images efficiently.

2) *Discriminative-based methods*: The discriminative-based method has also been utilized to overcome the challenges related to appearance changes during visual tracking. The discriminative method are called tracking-by-detection. The mechanism for discriminative-based methods is a separate set of features that are extracted such that they distinguish the target from the background image. A binary separation approach is used for target identification from the background in successive frames. Various studies have been conducted based on discriminative-based methods, such as the discriminative learning method based on graph embedding proposed by Zhang et al. [37]. Fan et al. [38] presented an approach of discriminative region attention that describes the target from the background in terms of spatial features. Their proposed method aimed to overcome the spatial distraction in the visual appearance changes challenging. Tang et al. [39] proposed a robust visual tracking method, DRLTracker, based on a discriminative ranking list approach. DRLTracker utilizes the ranking lists and two-scale features to generate a model of the target and recognize it from the background based on the ranking lists of generated patches. However, the proposed method is limited to two-scales DRLTracker and suffers from high processing time.

This study investigates an enhanced discriminative-based appearance modeling approach to overcome the noise shortcoming and appearance variations difficulty. This study uses appearance modeling instead of discriminative-based appearance modeling term in this paper. The appearance

modeling approach details are discussed in the Material and Methods section.

Finally, the core of this paper is the proposal of visual object tracking based on a combination of spatiotemporal saliency appearance modeling and sample feature-based approaches. The proposed method is based on a tracking-by-detection approach to provide a robust visual tracking system in appearance variation conditions. Correspondingly, a semi-automatic trigger-based algorithm is proposed to handle the phases' operation. Furthermore, an automatic algorithm is proposed to detect the region saliencies in a parallel implementation that leads to low processing time. Consequently, the temporal and spatial saliencies are integrated to generate the final saliency and sample features. Our contributions can be summarized as follows,

- A visual tracking method based on spatiotemporal saliency-based appearance modeling (SSAM) and sample feature-based target detection (SFTD) to preserve the visual target tracking robustly under appearance variation conditions, unpredictable motion, and low processing time. The proposed method is efficient in both camera and target moving platforms.
- Develop a novel algorithm for switching automatically between phases to handle their operation based on trigger activation.
- An algorithm is proposed for multiple target detection based on dynamic multithreading implementation of SLIC segmentation algorithm.

The remainder of this paper is organized as follows. Material and Methods section discusses the proposed framework and details of the material and methods. The results and discussion section presents our experimental and performance analysis. Finally, the conclusion section concludes this study.

### III. MATERIAL AND METHODS

This section presents an overview and the details of the proposed approach. The underlying goal of the proposed approach is to take advantage of saliency values and appearance modeling in an efficient manner for target detection.

In this study, the proposed method consists of two main phases, spatiotemporal saliency appearance modeling (SSAM) and sample feature-based target detection (SFTD). To handle the phases' operation, a semi-automatic trigger-based algorithm is proposed to switch between the two phases; a phase operation is started when that phase receives a trigger activation. For example, when the saliency-slot time is reached, the SSAM phase activates a trigger to switch to the SFTD phase. The proposed method defines the saliency slot to activate the trigger. The SFTD phase activates a trigger when it cannot detect any objects. The overall architecture of the proposed framework is shown in Fig. 1. The details of the proposed approach are presented in the following sections.

#### A. Spatiotemporal Saliency Appearance Modeling (SSAM) Phase

This phase involves three stages, temporal saliency and localization detection, spatial and final saliency detection, and, finally, sample feature generation and target detection stages.

1) *Temporal saliency and localization detection (TSLD) stage:* This stage consists of temporal saliency detection and localization modules. In order to extract the moving target, salient regions are extracted using motion cues detection. For motion cue detection, temporal saliency is investigated with the following details.

2) *Temporal saliency detection module:* The purpose of temporal saliency is for attention region (ARs) extraction and coarse segmentation. The extracted attentional regions are called Candidate Motion Regions (CMRs). To extract the CMRs, we propose the following steps,

Frame differencing is used for temporal saliency detection. Frame differencing is utilized to identify moving objects in consecutive frames. This technique employs the image subtraction operator, which takes two images (or frames) as input and produces the output [42].

Image Enhancement. Morphological operations are generally applied for image enhancement [43]. This proposed method uses these operators, which are dilation, erosion, and opening and closing; the morphological operators are inspired by [44, 45] with adapted structuring elements parameters.

3) *Localization module:* Once the temporal saliency detects the moving target region and enhances the CMRs, a localization module is applied to localize the extracted CMRs based on connected components and blob identification methods [47]. This module involves the following steps.

Thresholding. Thresholding assists us in reducing the number of false positives and avoiding missed valid objects. The thresholding is based on the variation of intensity consideration between the object pixels and the background pixels, as inspired by [48]. Setting a determined value to identify those pixels to implement the thresholding. In this matter, THRESH\_OTSU is used to determine the optimal threshold value using Otsu's algorithm [49].

Edge segmentation. Canny edge segmentation is then run on the binarized image for further improvement of the extracted region.

*Blob Identification.* After all the region's edges are extracted, we need to detect the blobs (connected components). To identify the blobs, active contour features, such as the one proposed by [50], are utilized to detect regions of interest and localize them. In this paper, we also use active contour features to detect the contours of regions. The detected contour features from CMRs regions are used for connected component detection, blob area, and bounding box determination. Moreover, removing unwanted blobs with a pixel area smaller than  $A_{low}$  or a bounding box with dimensions larger than  $B_{max}$ .

Candidate Mask Generation. In this step, geometrical features are extracted from the regions to recognize their location in each frame. Extracting  $X_{pos}$ ,  $Y_{pos}$  as the centroid of

each object based on moment features, as described in the spatial saliency detection module section later. Furthermore, we experimentally found the appropriate width and height values to generate the candidate mask (CM), extracting the regions based on the region of interest (ROI) function. Fig. 2 shows the generated candidate mask by the proposed temporal saliency and localization stage.

4) *Spatial saliency detection (SSD) stage*: The result of the TSLD stage is one or multiple CM regions, which are ARs. However, as shown in Fig. 2, some regions are incorrectly extracted that are unrelated to a target object or include useless regions. The spatial saliency detection (SSD) is used to overcome this fault detection. Furthermore, because candidate masks are compact and informative, we also investigate SSD to extract the saliency over them to provide further information and generate sample features. Our proposed SSD algorithm is based on integrating the proposed methods in [2, 29] with several modifications in feature extraction, feature representation, and spatial distribution measurement to improve the efficiency of spatial saliency extraction.

In brief, the input image is first decomposed into perceptually homogeneous segments as patches based on a SLIC superpixel algorithm presented in [51]. Second, we extract visual features, including color and moment, to measure the uniqueness and compactness of the spatial distribution. Finally, the temporal and spatial information are integrated to generate a final saliency map named spatiotemporal saliency.

However, since the SLIC is time-consuming for spatial saliency, we implement spatial saliency detection via parallel processing based on multi-threading programming. The use of multi-threading assists us in processing all CM regions in parallel. It can impressively decrease the overall processing time of the SSD stage. In this regard, each thread captures a candidate mask and performs the following processes to determine the spatial saliency and sample feature generation. For instance, if the result of the TSLD stage includes four objects, we assign each object to a thread, totaling four. OpenMP multi-threading is used as a tool to implement our spatial saliency detection algorithm. The steps for the SSD stage are as follows.

5) *Patch generation module*: Super-pixels segmentation as an effective region-based analysis algorithm is increasingly investigated by many researchers in computer vision communities [36,52]. As proposed in [51], this study uses a SLIC algorithm to segment the CM regions into homogeneous regions. Fig. 3 shows patch generation for a moving object using the SLIC superpixel algorithm.

6) *Spatial saliency module*: In this study, we use spatial uniqueness and compactness to compute the spatial saliency detection inspired by Perazzi et al. [53]. Moreover, we take advantage of other features, such as image moments and different metrics, to improve efficiency. In our method, we investigate pixel intensity for dissimilarity measurement of a patch compared to other regions. Compactness spatial distribution also contributes to detecting salient objects based

on image moments for uniqueness measurement. Details of the proposed spatial saliency detection are explained in the following.

Spatial uniqueness measurement. Similar to [54], each region's color similarity with other regions is measured. However, in [54], they implemented the color feature in a static image. In contrast, we investigate saliency detection in a dynamic environment. Furthermore, as reported in [55], Earth Mover's Distance (EMD) yielded excellent retrieval performance for the small sample size; we also use the EMD distance metric instead of Euclidean.

Spatial compactness measurement. Because the salient patches are spatially compact, the pixels with high saliency values are also expected to be spatially close [56]. Spatial moments are efficient and powerful in describing spatial distribution and compactness. In this study, we investigate spatial moments to estimate spatial compactness. Our work employs first- and second-order spatial moments.

7) *Final saliency map generation*: Generally, it is necessary to collaborate the temporal and spatial saliencies in a meaningful way to produce the final spatiotemporal saliency maps [10]. Therefore, the temporal and spatial information are integrated to generate a final saliency map named spatiotemporal saliency.

8) *Sample generation and target detection (SGDT) stage*: As shown in Fig. 1, the sample feature generation and target detection stage involve feature extraction, sample feature generation, and target detection. According to the feature extraction step and the result of the SSD stage, we collect appropriate features, such as color contrast and region compactness. As mentioned previously, these features are dynamically updated per frame and normalized, generating the sample features. Based on the sample features, we can detect the target.

#### B. Sample Feature-based Target Detection (SFTD) Phase

A trigger is activated upon the sample features being generated, and the sample features are transferred from the SSAM phase to this phase. The advantage of this phase is that it covers both moving and non-moving object detection conditions to detect objects with uncontrollable and unpredictable target movement conditions and overcome the difficulty of frame difference. The steps for this phase are mostly similar to the previous operation's steps, i.e., frame differencing, Image Enhancement, Feature Matching, Object Segmentation, and, finally, Target Detection.

## IV. RESULTS AND DISCUSSION

This section presents the implementation details and experimental results. Additionally, we compare the results with existing methods based on qualitative and quantitative performance evaluations to test and evaluate the proposed method. The qualitative analysis presents the image results from the proposed and other methods. In contrast, quantitative analysis involves precision and recall calculation and processing time. To validate the efficacy of the proposed method, the experiment was conducted on the VIVID public

dataset [57]. The VIVID dataset was collected at Eglin during DARPA VIVID and involves aerial images in video sequences. Several videos have been collected in VIVID, of which we use the EgTest01 and EgTest02 videos. The EgTest01 video involves moving cars that pass each other, with an image size of 640\*480 pixels and 1800 frames, whereas the EgTest02 video involves 1300 frames with two sets of three civilian vehicles passing each other on a runway.

A. Qualitative Analysis

Qualitative analysis is implemented to demonstrate the result of each phase and compare the proposed method with others. Fig. 4 shows the results of the qualitative analysis. The saliency-based methods considered for comparison are Itti [58], MD [21], GBVS [59], and SD [2]. Fig. 5 shows the comparison of the proposed method with other existing methods. The first row is the original raw images (Raw), the second, third and fourth are the results for the TSLD, SSD, and MOED phases, respectively, and the final row represents the feature-based object detection phase.

In the following sections, quantitative analysis for precision, recall, and f-measure calculation is discussed.

B. Precision and Recall Measurement

Similar to Refs. [2, 31, 60], precision and recall measures are used to evaluate the performance of the proposed method. In our evaluation, the target is the exciting object, whether moving or not. To measure the precision, recall, and f-measure, we need to define the following terms,

- True Positive: Detected salient regions that correspond to a target.
- False Positive: Detected salient regions that do not correspond to a target.
- False Negative: No detection of salient regions where there is, in fact, a target,

$$\text{Precision} = \left( \sum_{i=1}^n \frac{TP}{TP + FP} \right) \times 100 \%$$

$$\text{Recall} = \left( \sum_{i=1}^n \frac{TP}{TP + FN} \right) \times 100 \%$$

$$F_1 - \text{score} = 2 \times \left( \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \right)$$

Details of the measurements for TP, FP, FN, precision, recall, and F1-score are presented for the proposed method in Table I.

The precision and recall rates of different numbers of frames are illustrated in Fig. 6. As shown in Fig. 6 and 7, precision and recall rates are increased when the number of frames is increased.

Furthermore, to validate the proposed method, we compare our model with state-of-the-art visual object tracking methods, such as the FMD [2], DMM [60], HSC [10], RD [2], and SD [2]. The comparison was conducted based on precision, recall (PR), and F1-score. Table II and Fig. 8 show the comparison results. Based on the obtained experimental results, we show that the proposed approach can be effectively employed for the extraction of moving objects.

C. Processing Time

Our experiment was implemented in Visual Studio and performed on a Windows 8 platform with an Intel 2.6 GHz CPU and 4 GB of Memory. The processing time is measured based on wall-clock time computation because, when measuring the performance of parallel programs, the wall-clock time needs to be considered, then using the tick\_count class, which is located in `tbb/tick_count.h`. A tick\_count is an absolute timestamp. The average processing time for the proposed method is approximately 78 and 24 milliseconds for SSAM and SFTD, respectively, which is suitable for near-real-time visual tracking applications.

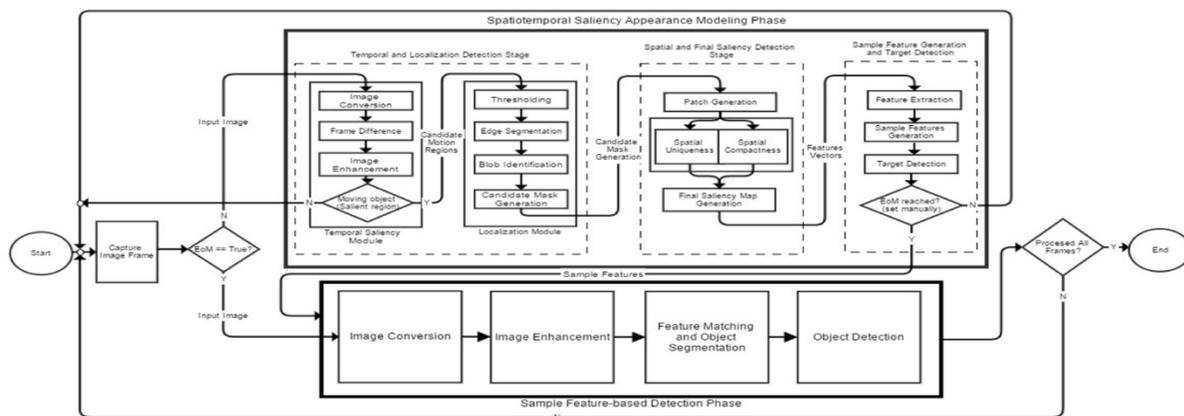


Fig. 1. Our proposed framework for visual tracking system

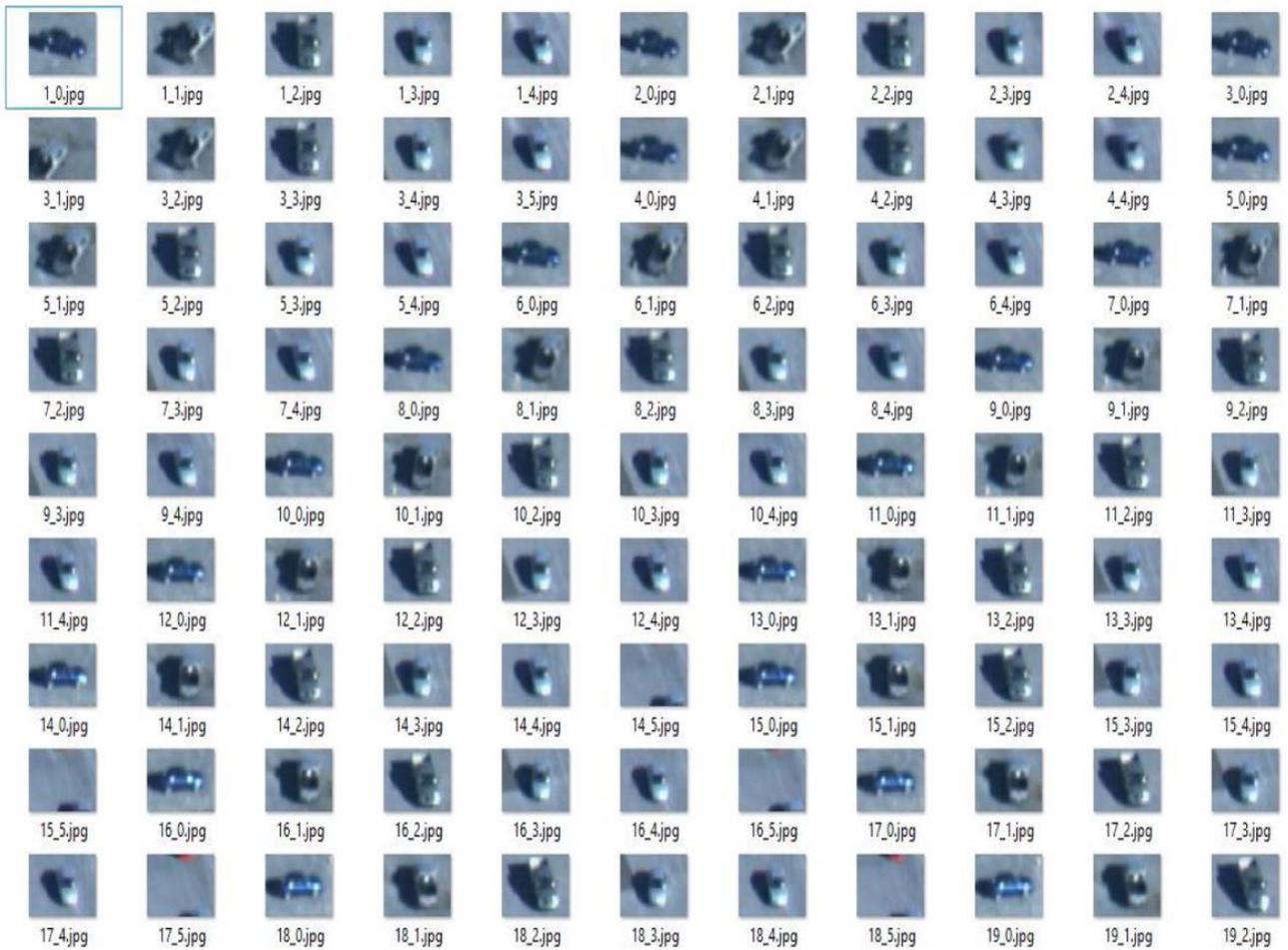


Fig. 2. Candidate mask generation images

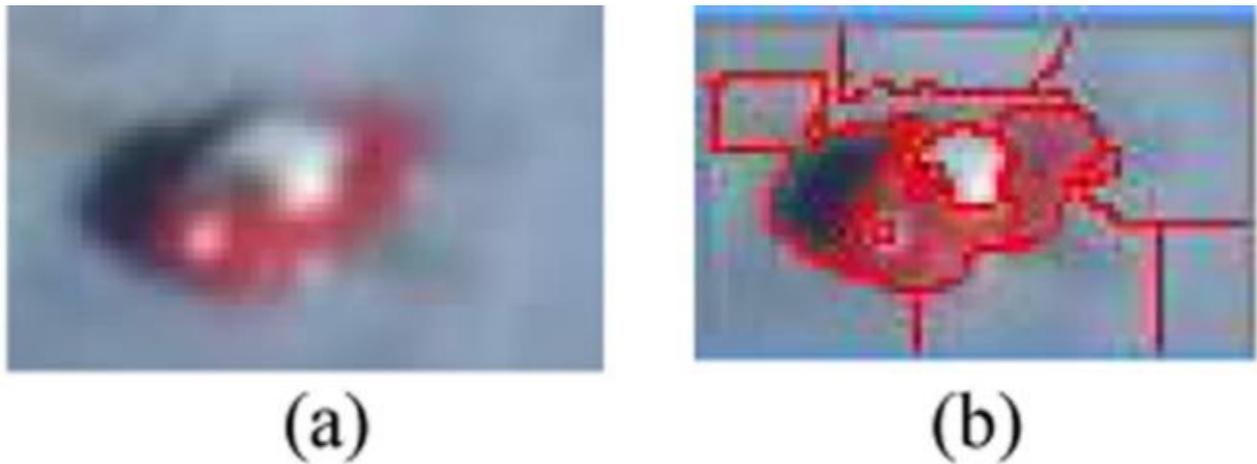


Fig. 3. Patch generation for a moving object using a parallel SLIC superpixel algorithm. (a) An original candidate mask, (b) generated patches

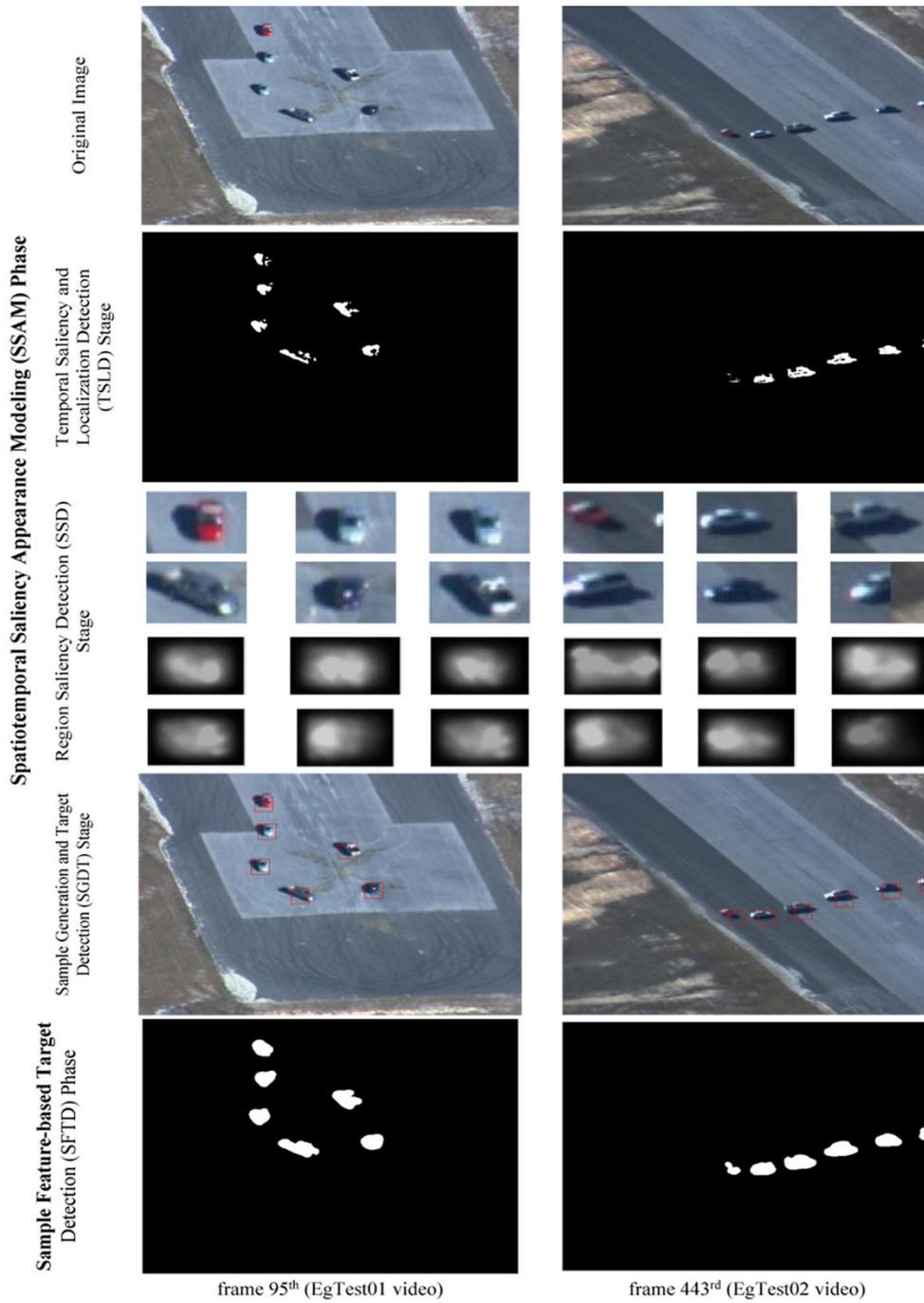


Fig. 4. Image results for each phase and stage of the proposed method

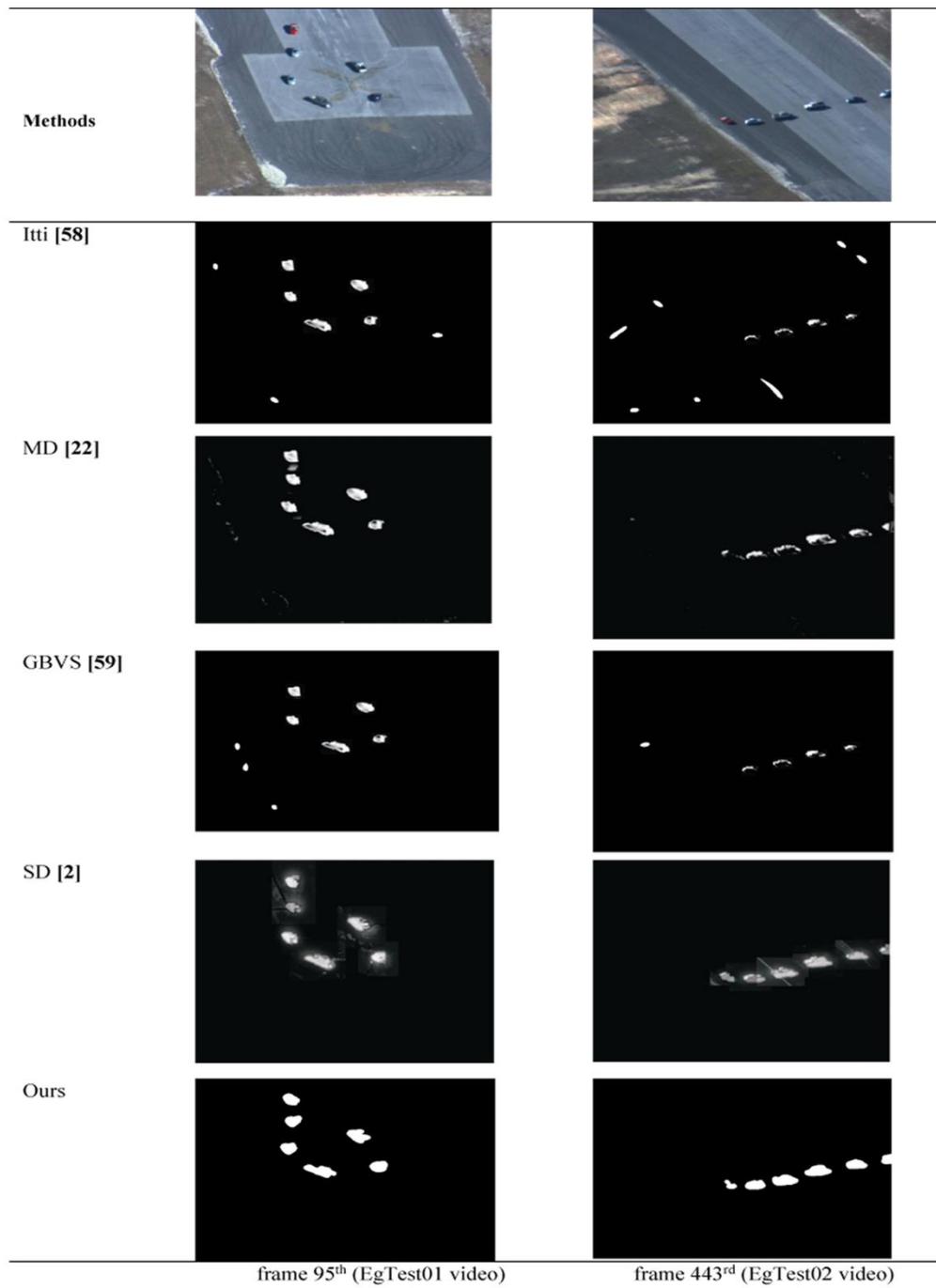


Fig. 5. Comparison of the proposed method with four state-of-the-art saliency methods

TABLE I. DETAILS OF THE MEASUREMENTS OF THE TRUE POSITIVE, FALSE POSITIVE, FALSE NEGATIVE, PRECISION, AND RECALL RATES

Dataset	Number of Frames	TP	FP	FN	Precision	Recall
					(%)	(%)
EgTest01	50	32	12	6	0.73	0.84
	450	307	97	46	0.76	0.87
	1150	846	214	93	0.80	0.90
	1800	1381	298	121	0.82	0.92

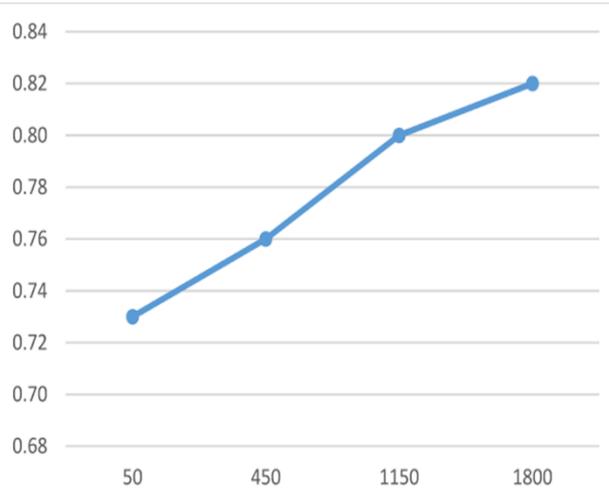


Fig. 6. Precision metric comparison for different numbers of frames

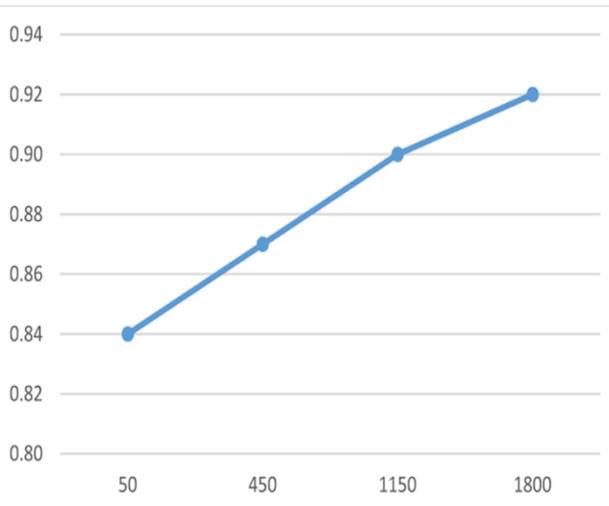


Fig. 7. Comparison of different numbers of frames based on the recall measures

TABLE II. COMPARISON OF VISUAL TRACKING METHODS AND THE PROPOSED METHOD

Method	Recall (%)	Precision (%)	F <sub>1</sub> -score (%)
FMD	0.49	0.34	0.40
DMM	0.68	0.48	0.56
HSC	0.69	0.51	0.59
RD	0.74	0.69	0.71
SD	0.79	0.48	0.60
Ours	0.82	0.73	0.77

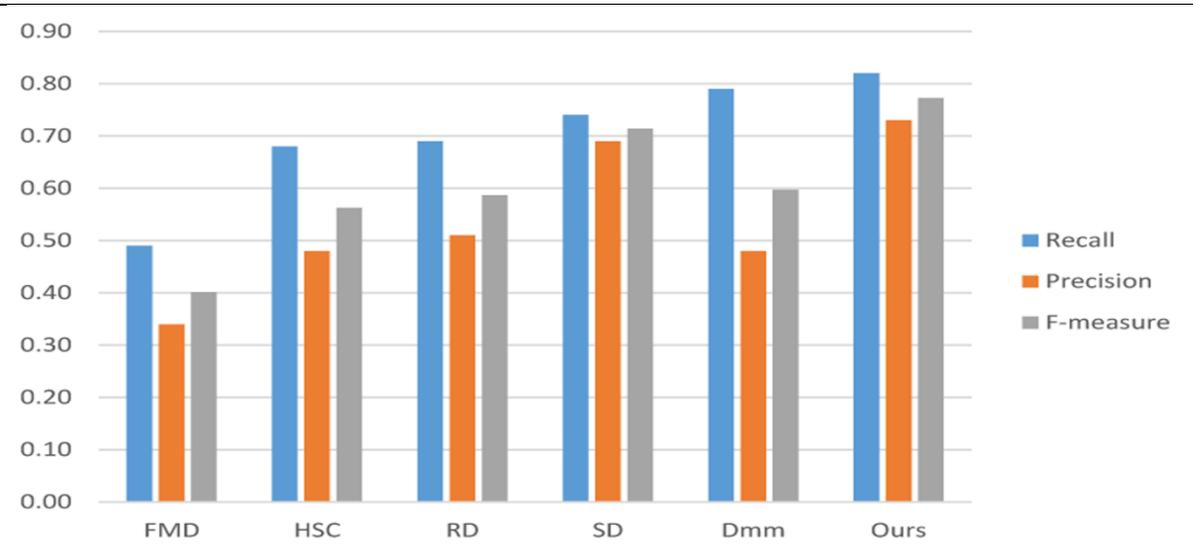


Fig. 8. Precision, recall and F-measure for visual tracking methods and our method

## V. CONCLUSION

This paper addresses the significant problems facing visual tracking, such as appearance variations and unpredictable moving targets, for aerial images. The proposed method uses spatial and temporal saliencies to address these challenges by adopting biologically inspired approaches to detect the attentional regions (ARs). Furthermore, a biologically inspired approach integrated with an appearance modeling-based approach is investigated to overcome visual tracking challenges. In this regard, the proposed method consists of two main phases, spatiotemporal saliency-based appearance modeling (SSAM) and sample feature-based target detection (SFTD). The proposed method uses a tracking-by-detection approach to provide a robust visual tracking system under appearance variation conditions. Correspondingly, a semi-automatic trigger-based algorithm is proposed to handle the phases' operation, and a discriminative-based method is utilized for appearance modeling. In the spatiotemporal saliency phase, temporal saliency is used to extract the attentional regions (ARs) and coarse segmentation. Spatial saliency is utilized to obtain the object's appearance details in ARs regions. By combining temporal and spatial saliencies, we can obtain refined detection results and track the target. During

the spatial saliency detection, prominent features are collected, and a sample feature is generated to describe the target.

Consequently, a target detection process is performed to recognize the target in images. Experiments were conducted on the VIVID dataset. Moreover, the proposed method compared with other state-of-the-art methods. The analyses demonstrate that the proposed method is superior to most state-of-the-art methods and presents an effective visual tracking method which is robust in appearance variation difficulties.

Future works can be conducted to address other difficulties and challenges in visual tracking, such as when complicated backgrounds or backgrounds with partial and/or full occlusion are present.

## ACKNOWLEDGMENT

The authors would like to appreciate Assoc. Prof. Dr. Anton Satria Prabuwono, Dr. Ang Mei Choo, and Teck Loon Lim for helpful advice and suggestions. We also thank all the anonymous reviewers for their comments which assisted us in improving the quality of this paper.

REFERENCES

- [1] K. Zhang, L. Zhang, Q. Liu, D. Zhang, and M.-H. Yang (2014), Fast Visual Tracking via Dense Spatio-temporal Context Learning, In: Computer Vision–ECCV 2014: Springer, pp. 127-141.
- [2] H. Shen, S. Li, C. Zhu, H. Chang, and J. Zhang (2013), Moving object detection in aerial video based on spatiotemporal saliency, Chinese Journal of Aeronautics, vol. 26, pp. 1211-1217.
- [3] F. Chen, Q. Wang, S. Wang, W. Zhang, and W. Xu (2011), Object tracking via appearance modeling and sparse representation, Image and Vision Computing, vol. 29, pp. 787-796.
- [4] S. Zhang, H. Yao, H. Zhou, X. Sun, and S. Liu (2013), Robust visual tracking based on online learning sparse representation, Neurocomputing, vol. 100, pp. 31-40.
- [5] H. Yang, L. Shao, F. Zheng, L. Wang, and Z. Song (2011), Recent advances and trends in visual tracking: A review, Neurocomputing, vol. 74, pp. 3823-3831.
- [6] M. Yang, J. Yuan, and Y. Wu, Spatial selection for attentional visual tracking, in Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on, 2007, pp. 1-8.
- [7] C. Siagian and L. Itti (2007), Rapid biologically-inspired scene classification using features shared with visual attention, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 29, pp. 300-312.
- [8] Y. Kashiwase, K. Matsumiya, I. Kuriki, and S. Shioiri (2013), Temporal Dynamics of Visual Attention Measured with Event-Related Potentials, PLoS one, vol. 8, p. e70922.
- [9] Y. Zhai and M. Shah, Visual attention detection in video sequences using spatiotemporal cues, in Proceedings of the 14th annual ACM international conference on Multimedia, 2006, pp. 815-824.
- [10] C. Li, J. Xue, N. Zheng, X. Lan, and Z. Tian (2013), Spatio-temporal saliency perception via hypercomplex frequency spectral contrast, Sensors, vol. 13, pp. 3409-3431.
- [11] L. Itti, Models of bottom-up and top-down visual attention, California Institute of Technology, 2000.
- [12] F. Göschl, A. K. Engel, and U. Fries (2014), Attention Modulates Visual-Tactile Interaction in Spatial Pattern Matching, PLoS one, vol. 9, p. e106896.
- [13] V. Mahadevan and N. Vasconcelos (2013), Biologically Inspired Object Tracking Using Center-Surround Saliency Mechanisms, Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 35, pp. 541-554.
- [14] A. Borji and L. Itti (2013), State-of-the-art in visual attention modeling, Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 35, pp. 185-207.
- [15] W. Kim, C. Jung, and C. Kim (2011), Spatiotemporal saliency detection and its applications in static and dynamic scenes, Circuits and Systems for Video Technology, IEEE Transactions on, vol. 21, pp. 446-456.
- [16] D. Kerzel, J. Schönhammer, N. Burra, S. Born, and D. Souto (2011), Saliency changes appearance, PLoS one, vol. 6, p. e28292.
- [17] Y. Zhang, Z. Mao, J. Li, and Q. Tian (2014), Salient Region Detection for Complex Background Images Using Integrated Features, Information Sciences,
- [18] H. R. Tavakoli, E. Rahtu, J. Heikkil, and #228 (2013), Temporal saliency for fast motion detection, presented at the Proceedings of the 11th international conference on Computer Vision - Volume Part I, Daejeon, Korea.
- [19] T. Crivelli, P. Boutheymy, B. Cernuschi-Frías, and J.-f. Yao (2011), Simultaneous motion detection and background reconstruction with a conditional mixed-state markov random field, International journal of computer vision, vol. 94, pp. 295-316.
- [20] O. Barnich and M. Van Droogenbroeck (2011), ViBe: A universal background subtraction algorithm for video sequences, Image Processing, IEEE Transactions on, vol. 20, pp. 1709-1724.
- [21] Z. Yin and R. Collins, Moving object localization in thermal imagery by forward-backward MHI, in Computer Vision and Pattern Recognition Workshop, 2006. CVPRW'06. Conference on, 2006, pp. 133-133.
- [22] C. Benedek, T. Szirányi, Z. Kato, and J. Zerubia (2009), Detection of object motion regions in aerial image pairs with a multilayer markovian model, Image Processing, IEEE Transactions on, vol. 18, pp. 2303-2315.
- [23] G. Medioni, I. Cohen, F. Brémond, S. Hongeng, and R. Nevatia (2001), Event detection and analysis from video streams, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 23, pp. 873-889.
- [24] K. K. Ng and E. J. Delp, Background subtraction using a pixel-wise adaptive learning rate for object tracking initialization, in IS&T/SPIE Electronic Imaging, 2011, pp. 78820I-78820I-9.
- [25] K. K. Ng and E. J. Delp, Object tracking initialization using automatic moving object detection, in IS&T/SPIE Electronic Imaging, 2010, pp. 75430M-75430M-12.
- [26] H. J. Min, Multi-Robot Formation and Cooperation Using Visual Tracking, 2013.
- [27] X. Shen and Y. Wu, A unified approach to salient object detection via low rank matrix recovery, in Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, 2012, pp. 853-860.
- [28] E. Erdem and A. Erdem (2013), Visual saliency estimation by nonlinearly integrating features using region covariances, Journal of vision, vol. 13, p. 11.
- [29] W. Wang, D. Cai, X. Xu, and A. Wee-Chung Liew (2014), Visual saliency detection based on region descriptors and prior knowledge, Signal Processing: Image Communication, vol. 29, pp. 424-433.
- [30] M.-M. Cheng, G.-X. Zhang, N. J. Mitra, X. Huang, and S.-M. Hu, Global contrast based salient region detection, in Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, 2011, pp. 409-416.
- [31] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, Frequency-tuned salient region detection, in Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, 2009, pp. 1597-1604.
- [32] J. Harel, C. Koch, and P. Perona, Graph-based visual saliency, in Advances in neural information processing systems, 2006, pp. 545-552.
- [33] K.-C. Lee and D. Kriegman, Online learning of probabilistic appearance manifolds for video-based recognition and tracking, in Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on, 2005, pp. 852-859.
- [34] Y. Wu, J. Cheng, J. Wang, H. Lu, J. Wang, H. Ling, et al. (2012), Real-time probabilistic covariance tracking with efficient model update, IEEE Transactions on Image Processing, vol. 21, pp. 2824-2837.
- [35] T. Bai and Y. F. Li (2012), Robust visual tracking with structured sparse representation appearance model, Pattern Recognition, vol. 45, pp. 2390-2404.
- [36] B. Zhong, Y. Chen, Y. Shen, Y. Chen, Z. Cui, R. Ji, et al. (2014), Robust tracking via patch-based appearance model and local background estimation, Neurocomputing, vol. 123, pp. 344-353.
- [37] X. Zhang, W. Hu, S. Maybank, and X. Li, Graph based discriminative learning for robust and efficient object tracking, in Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on, 2007, pp. 1-8.
- [38] J. Fan, Y. Wu, and S. Dai (2010), Discriminative spatial attention for robust tracking, In: Computer Vision–ECCV 2010: Springer, pp. 480-493.
- [39] M. Tang and X. Peng (2012), Robust tracking with discriminative ranking lists, IEEE Transactions on Image Processing, vol. 21, pp. 3273-3281.
- [40] C. Huang, Q. Liu, and S. Yu (2011), Regions of interest extraction from color image based on visual saliency, The Journal of Supercomputing, vol. 58, pp. 20-33.
- [41] W. Chen, Y. Q. Shi, and G. Xuan, Identifying computer graphics using HSV color model and statistical moments of characteristic functions, in Multimedia and Expo, 2007 IEEE International Conference on, 2007, pp. 1123-1126.
- [42] R. C. Gonzalez and R. E. Woods, Digital image processing, ed: Prentice hall Upper Saddle River, NJ., 2002.
- [43] K. Sreedhar and B. Panlal (2012), Enhancement of images using morphological transformation, arXiv preprint arXiv:1203.2514,

- [44] E. R. Dougherty, R. A. Lotufo, and T. I. S. f. O. E. SPIE, Hands-on morphological image processing vol. 71: SPIE press Bellingham, 2003.
- [45] J. Serra (1986), Introduction to mathematical morphology, Computer vision, graphics, and image processing, vol. 35, pp. 283-305.
- [46] X. Bai, F. Zhou, and B. Xue (2012), Image enhancement using multi scale image features extracted by top-hat transform, Optics & Laser Technology, vol. 44, pp. 328-336.
- [47] A. Das, M. Diu, N. Mathew, C. Scharfenberger, J. Servos, A. Wong, et al. (2014), Mapping, Planning, and Sample Detection Strategies for Autonomous Exploration, Journal of Field Robotics, vol. 31, pp. 75-106.
- [48] W. OpenCV. (2014). Basic Thresholding Operations. Available: [http://docs.opencv.org/3.0-alpha/doc/py\\_tutorials/py\\_imgproc/py\\_thresholding/py\\_thresholding.html?highlight=adaptive%20thresholding](http://docs.opencv.org/3.0-alpha/doc/py_tutorials/py_imgproc/py_thresholding/py_thresholding.html?highlight=adaptive%20thresholding)
- [49] W. OpenCV. (2014). Miscellaneous Image Transformations. Available: [http://docs.opencv.org/modules/imgproc/doc/miscellaneous\\_transformations.html?highlight=threshold#threshold](http://docs.opencv.org/modules/imgproc/doc/miscellaneous_transformations.html?highlight=threshold#threshold)
- [50] S. Sclaroff and J. Isidoro (2003), Active blobs: region-based, deformable appearance models, Computer Vision and Image Understanding, vol. 89, pp. 197-225.
- [51] C. Y. Ren and I. Reid (2011), gSLIC: a real-time implementation of SLIC superpixel segmentation, University of Oxford, Department of Engineering, Technical Report,
- [52] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk (2012), SLIC superpixels compared to state-of-the-art superpixel methods, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 34, pp. 2274-2282.
- [53] F. Perazzi, P. Krahenbuhl, Y. Pritch, and A. Hornung, Saliency filters: Contrast based filtering for salient region detection, in Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, 2012, pp. 733-740.
- [54] L. Shuhua, L. Zhi, L. Lina, Z. Xuemei, and O. Le Meur, Efficient saliency detection using regional color and spatial information, in Visual Information Processing (EUVIP), 2013 4th European Workshop on, 2013, pp. 184-189.
- [55] Y. Rubner, C. Tomasi, and L. J. Guibas (2000), The earth mover's distance as a metric for image retrieval, International Journal of Computer Vision, vol. 40, pp. 99-121.
- [56] Z. Chi and W. Weiqiang, Object-level saliency detection based on spatial compactness assumption, in Image Processing (ICIP), 2013 20th IEEE International Conference on, 2013, pp. 2475-2479.
- [57] S. o. C. S. Robotics Institute, Carnegie Mellon University. . (2013). VIVID Tracking Evaluation Web Site. Available: <http://vision.cse.psu.edu/data/vividEval/datasets/PETS2005/EgTest01/index.html>
- [58] L. Itti, C. Koch, and E. Niebur (1998), A model of saliency-based visual attention for rapid scene analysis, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 20, pp. 1254-1259.
- [59] J. Harel, C. Koch, and P. Perona, Graph-based visual saliency, in Advances in neural information processing systems, 2007, pp. 545-552.
- [60] F. M. S. Saif, A. S. Prabuwno, and Z. R. Mahayuddin (2014), Moving Object Detection Using Dynamic Motion Modelling from UAV Aerial Images, The Scientific World Journal, vol. 2014, p. 12.

# A Study on Distance Personalized English Teaching Based on Deep Directed Graph Knowledge Tracking Model

Lianmei Deng

Department of Economics and Trade, Chongqing College of Finance and Economics, Chongqing, 402160, China

**Abstract**—Despite the continuous development of online education models, the effectiveness of online distance education has never been able to meet people's expectations due to individual difference of learners. How to personalize teaching in a targeted manner and stimulate learners' independent learning ability has become a key issue. In this study, the multidimensional features of the learning process are mined with the help of the BORUTA feature selection model, and the DKVMN-BORUTA model incorporating multidimensional features is established. This optimized deep knowledge tracking method is combined with graph structure rules. Then, an intelligent knowledge recommendation algorithm based on reinforcement learning is used to construct a fusion approach-based model for distanced personalized teaching and learning of English. The results show that the research proposed fused deep-directed graph knowledge tracking with graph structure rules for remote personalized English teaching model has the lowest AUC value of 0.893 and the highest AUC value of 0.921 on each dataset. The prediction accuracy of the research model is 94.3% and the F1 score is 0.92, which is the highest among the studied models, indicating that the proposed model has a strong performance. The fusion model proposed in the study has a higher accuracy rate of knowledge personalization recommendation than the traditional deep knowledge tracking model, and it can help learners save revision time effectively and improve their overall English performance.

**Keywords**—Distanced personalized English teaching model; knowledge tracking; deep learning; graph structure rules; DKVMN

## I. INTRODUCTION

With the booming development of internet information technology, online educational resources are shared and disseminated nationwide [1]. Online distanced English teaching is particularly outstanding in terms of resource sharing and students are very fond of being taught by foreign teachers via online interactive communication. However, as the online teaching method is quite different from the traditional face-to-face teaching method, it put forwards higher requirements for students' independent learning ability and self-motivation. Therefore, a more scientific and effective way is required to track students' status in the classroom, attract students' attention, achieve knowledge tracking, and improve the quality of distanced English learning [2-3]. In previous studies, some scholars have proposed to apply deep learning to the prediction of knowledge and to build Deep Knowledge Tracing (DKT) models by combining Long Short-Term Networks (LSTM) with history department sequences [4-5].

However, the DKT model is not very accurate in analyzing students' learning data and predicting knowledge, and its performance is weakly stable. At present, the improvement of distance personalized teaching quality mainly depends on the knowledge tracking of learners. To accurately analyze the knowledge state of learners, many technologies have been applied in the field of knowledge tracking. Among them, the most advanced and widely used ones mainly include two types, namely, the knowledge tracking method based on probability graph and the knowledge tracking method based on deep learning. However, the knowledge tracking method based on probability graph lacks the consideration of the relationship between learner forgetting factors and knowledge, so its accuracy is low. However, the knowledge tracking method based on deep learning is difficult to obtain the specific information of learners' knowledge state. For this reason, the study optimizes the DKT model based on data mining and improves the DKT model in terms of multidimensional features of the learning process and feature fusion. The optimized model is combined with graph structure rules, and then the distanced personalized English teaching model based on deep directed graph knowledge tracking is proposed. This paper first reviews the application of deep learning model in teaching, and then summarizes the existing defects of current methods through the study of previous methods. Then, the research methods are elaborated, and the deep learning knowledge tracking model is proposed, which is applied to English distance personalized education. Secondly, in the part of result analysis, the research algorithm is compared with the current more advanced knowledge tracking algorithm, and the effectiveness and accuracy of the research algorithm are verified. Finally, in the conclusion part, the advantages and disadvantages of the research algorithm are summarized, and the future research direction is proposed. It aims to precisely locate the learning status of learners and their mastery of knowledge points so as to achieve good results of distanced personalized English teaching.

With the rapid development of deep learning and neural network algorithms, the role of computer technology in the teaching and learning process has been explored, and scholars have begun to focus on related topics, and substantial progress has been achieved. GERVETT et al. applied deep learning DKT models to knowledge tracking and analyzed the conditions under which their deep models can make the most accurate predictions [6]. SAPOUNTZIA et al. analyzed the Bayesian Knowledge Tracing (BKT) model, the DKT model,

and the Dynamic Key-Value Memory Network (DKVMN) model from a technical and educational perspective and compared the modeling techniques, evaluation, and performance in terms of optimization of the three models, and the results showed superior performance of the DKVMN model [7]. Hassan et al. focused on student dropout risk prediction by using deep long and short-term memory models. Multidimensional data can help teachers predict students' learning status, and comparative testing results on real datasets show superior application performance of the model over logistic regression and artificial neural networks [8]. Mubarak et al. analyzed the role of video clickstream data in prediction of online learners' learning performance, and deep neural network algorithms incorporating implicit features were found to have excellent performance in real-world applications, which outperformed super vector machines and logistic regression methods [9]. Kaser et al. proposed a dynamic Bayesian knowledge tracing (DBKT) model that uses DBNs to combine multiple knowledge tracing (KC) in one model, and the results showed that the model had validity and strong contractual performance [10].

DEONOVICB et al. proposed a method to improve the prediction accuracy of knowledge tracking by combining probability maps and cognitive diagnostic techniques in order to improve the performance of knowledge tracking, and the experimental results showed that the method effectively improved the performance and prediction accuracy of knowledge tracking [11]. Yang et al. optimized the deep knowledge tracking technique in tracking students' knowledge states. They enriched the original deep knowledge tracking model by adding heterogeneous feature implicatures to determine the probability of students' correct answers in the exercises. The study evaluated the optimized model using two different education-related datasets, and the evaluation results indicated the superior performance of the model in the relevant domain [12]. Wang et al. designed the DKTS method using the feature relationship between topics and the linkage of knowledge points, and the experimental results showed that the method is effective in improving the service performance of the knowledge tracking method [13]. Vie et al. concluded that tracking the evolution of students' knowledge can help teachers' instructional optimization, and factor decomposer can be used as a regression and classification model. The findings show that the model can handle multi-feature high-dimensional student learning data in a large number of real datasets, and it has significant superiority over existing models [14]. This is the first work to integrate competency-based tracking into MOOC course recommendations. Extensive experiments on real-world datasets demonstrate that capacity tracking-enhanced course recommendations improve the effectiveness and interpretability of MOOCs. Tian et al. made an attempt to integrate capacity tracking into course recommendations for MOOCs. Experiments on a large amount of real-world data demonstrate that deep knowledge tracking can improve the validity and interpretability of MOOCs [15].

In summary, deep learning and neural network algorithms have been widely used in education and a large number of research findings have enriched current educational approaches, where deep knowledge tracking has also been widely used to track student knowledge evolution. However, there has been no reports investigating directed graphs as an input data source. Therefore, the research will focus on the construction of graph attention neural networks on this basis to obtain a student knowledge tracking model with better performance.

## II. DEEP DIRECTED GRAPH KNOWLEDGE TRACKING MODEL BASED ON PERSONALIZED ENGLISH DISTANCE TEACHING

### A. Design of a Deep Learning Knowledge Tracking Model based on Multidimensional Feature Fusion

Knowledge tracking automatically tracks changes in the state of knowledge of learners based on their historical learning trajectories and predicts learners' mastery of knowledge in the future process [16]. The study describes a learner's answer

interaction as  $R_i = \{s_i, q_i, r_i, b_i\}$ , where  $b_i$  is a sequence of learning behaviors. Considering the large number of feature dimensions and the need for knowledge tracking methods to mine the historical learning data collected by the learning platform, this research proposes a knowledge tracking method incorporating multidimensional features of the learning process, as shown in Fig. 1.

As shown in Fig. 1, the knowledge tracking framework consists of data composition, knowledge tracking optimization, and knowledge tracking model. Among them, the answer history and other data constitute the data framework, and the knowledge tracking optimization includes data mining and improvement methods for the deep knowledge tracking model to optimize the model and improve the prediction accuracy and interpretability of the model. Since deep knowledge tracking and other learning models have the problem of weak interpretability, some studies have pointed out that the interpretability can be divided into Ante-hoc and Post-hoc perspectives. Based on Post-hoc as the perspective for global interpretation, the multidimensional features that affect the learning results are explored from many features of the learning process, and then redundant features are removed to complete the input composition of the model, i.e., the BORUTA feature selection method [17]. Moreover, the knowledge tracking model mainly utilizes the Dynamic Key-Value Memory Networks (DKVMN) model to ensure the performance and interpretability of the method [18]. Since DKVMN has a strong expansion performance, the core of the research is the optimization of the deep knowledge tracking model incorporating multidimensional features, as shown in Fig. 2.

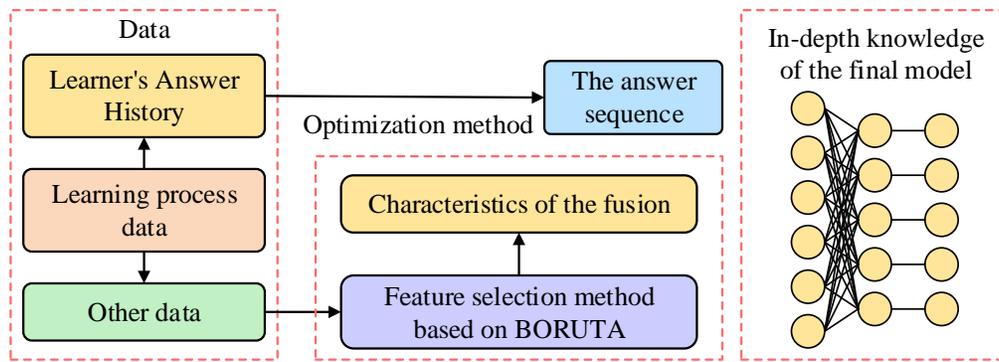


Fig. 1. Knowledge tracking framework based on learning process characteristics.

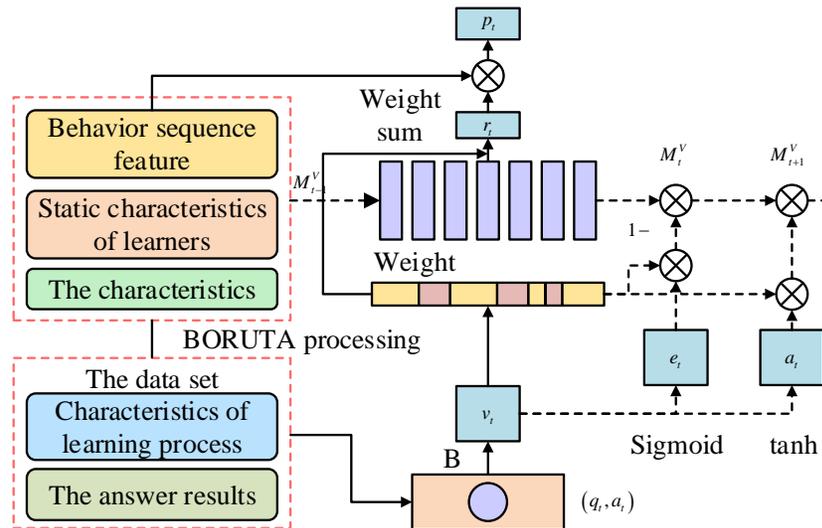


Fig. 2. Depth knowledge tracking model with multi-dimensional features.

As shown in Fig. 2, the processed data is used as the input of the model, and three networks, namely learning behavior sequence features, learner static features, and topic features, are built at the output location of this model, and then the optimization method is incorporated into the DKVMN framework, i.e., the deep knowledge tracking optimization method DKVMN-BORUTA is obtained. The study obtains the Embedding matrix of the topic by constructing its vector representation as in equation (1) for the knowledge concept weights[19].

$$w_t(i) = \text{soft max}(M_i^k g k_t) \quad (1)$$

Where, the problem  $q$  encountered by the learner at the moment  $t$  is first multiplied by the embedding matrix  $A$  that has been trained to obtain the embedding vector  $k_t$ . Then  $k_t$  is computed by  $\text{soft max}$  to obtain the attention weight vector  $w_t$ . The ease of expansion of DKVMN creates the conditions for the incorporation of multidimensional features in the learning process. Specifically, the DKVMN model receives learning records that will have some impact on the learner's feature vector  $f_t$ . Since  $f_t$  is obtained by fusing the read

vector  $r_t$  with the embedding vector  $k_t$ , it contains both information about the state of the learner's knowledge of the topic  $q_t$  and the embedding information of  $q_t$ . By processing the Fourier transform through the neural network, it is possible to infer the learner's ability on  $q_t$ . And the difficulty level of  $q_t$  can be obtained by passing  $k_t$  to the neural network.

Equation (2) is the formula for the read vector of  $r_t$

$$r_t = \sum_{i=1}^N w_t(i) M_i^v \quad (2)$$

Where,  $M_t^v$  represents the read value memory. The BORUTA features are processed through the embedding matrix to obtain the vector representation and then the embedding matrix is spliced with the topic content feature vector  $v_t$ . Eq. (3) is the formula for calculating the probability of a learner correctly answering the question  $q_t p_t$ .

$$p_t = \mathcal{O}\left(\tanh\left(W_o^T [r_t, m_t, l_t, v_t]\right) + b_{it}\right) \quad (3)$$

Where,  $l_t$  denotes the learner static features,  $m_t$  denotes the topic feature limit.  $b_{it}$  represents the conditioning factor, and  $\mathcal{O}$  is the Sigmoid activation function. Where  $b_{it} = \text{Sigmoid}|b_{it}|$ ,  $b_{it}$  is calculated by first mining the effective learning behavior through the learning behavior sequence feature vector. If the learner first watched the video and then discussed in the discussion forum and watched the learning video to answer the question, the learning behavior sequence of this learner is  $b_{it} = (2, 1)$ . The overall vector is a fusion of the BORUTA feature set with the current topic content features and the learner's knowledge state. The knowledge state matrix of the deep knowledge tracking optimization model is updated by inputting answer records  $(q_t, a_t)$  and  $w_t$  to collaboratively update  $M_t^v$ , and by erasing weights  $E$  and adding weights  $D$ , respectively, so as to obtain the memory erasure vector  $e$  and the memory addition vector  $a$ . The memory erasure vector is calculated by equation (4).

$$e_t = \text{Sigmoid}(W_e v + t) \quad (4)$$

Where,  $W_e$  represents the erasure weights and  $t$  represents the discretized value of the learner's time to deal with the problem. The memory addition vector is calculated by Eq. (5).

$$a_t = \tanh(W_a v_t + t) \quad (5)$$

Where,  $W_a$  stands for adding weights. The process of dynamic update of the learner's knowledge state is expressed by Eq. (6).

$$M_{t+1}^v = M_t^v [1 - w(i)e_t] [1 + w(i)a_t] \quad (6)$$

Where  $x_t = (q_t, a_t)$  represents the learner's answering behavior after the  $t$  moment and the value of the dynamic matrix  $M$ , which is transformed from  $M_t^v$  to  $M_{t+1}^v$ . The goal of model optimization is to minimize the difference between  $p_t$  and  $a_t$  with the minimum spread loss function. The model optimization is conducted using the momentum gradient descent method, as shown in equation (7).

$$L = -\sum_i [a_i \log p_i + (1 + a_i) \log (1 - p_i)] \quad (7)$$

In this study, the BORUTA feature selection model is used to explore the multi-dimensional features in the online learning process. Based on DKVMN model, the multi-dimensional feature network is constructed. Secondly, the deep knowledge tracking model with multi-dimensional learning features is designed. Finally, a deep knowledge tracking optimization model is constructed.

#### B. A Model Design for Distanced Personalized English Learning incorporating Deep Knowledge Tracking and Graph Structure Rules

In the design process of this research model, the deep knowledge tracking model was first introduced, allowing the construction of a directed graph of knowledge concepts for the probability values of the outcomes, and then directed graph was fed into the deep learning neural network as a data source, which in turn formed the attention neural network [20]. The probability value of the association relationship between concepts is actually the weight value of the network nodes. The structure of the deep directed knowledge tracking model is shown in Fig. 3.

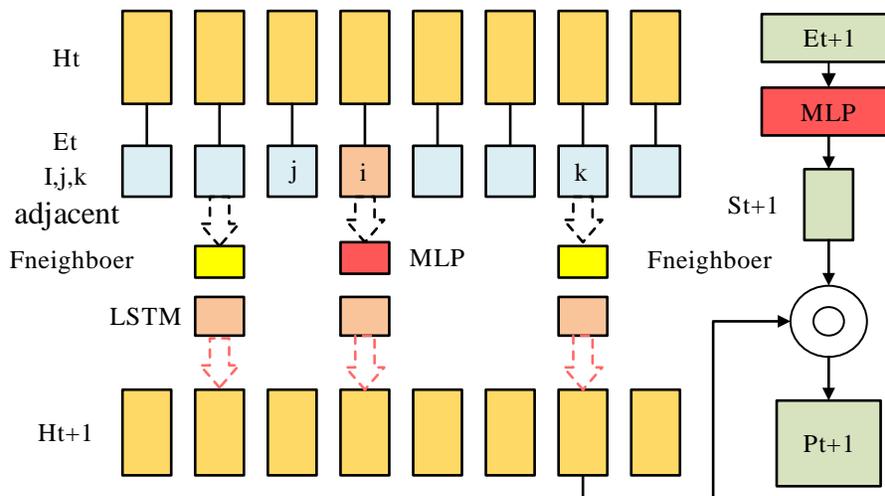


Fig. 3. Overall structure of knowledge tracking model integrating depth directed graph.

As shown in Fig. 3, the model first undergoes a Manifold Embedding popular embedding, setting up the memory linear matrix  $H$ , which holds mainly the basic state of the knowledge concept, and the corresponding  $N$  vector values of relatively independent concepts related to the problem. For the vectors,  $N d$  are used to represent their dimensions. The memory values in the matrix  $H$  at the time of  $t$  are  $H_t$ , and  $E$  is used to identify a matrix of processing result values after popular embedding. In the model processing, the extremely strong answer to the question at the  $t$  moment is processed as a one-bit valid code of length  $2N x_t$ , and the code is processed using the embedding matrix of  $2N * e$  to downscale the dimensionality to a vector of length  $e$ . For example, equation (8) is the vector matrix of.  $N * e$

$$E_k^t = \begin{cases} xtEx(k=i) \\ Ec(k)(k \neq i) \end{cases} \quad (8)$$

Where  $x_t$  represents the heat code and  $e_x$  represents the vector value. From Fig. 3, it can be seen that the model input value at the moment of  $t$  is  $(q_t, a_t)$ , and after that knowledge concept is defined as  $i$ , the directed graph connecting the adjacent nodes of  $i, j, k$  and the vector corresponding to  $i, j, k$  in the matrix of synchronous negative memory concepts  $h$  is established, as shown in Eq. (9).

$$H_t = [H_t, E_t] \quad (9)$$

Where the vector  $h_t^i$  is spliced with the vector  $e_t^i$ , processed by a fully connected neural network containing hidden and input layers, and then fed into the LSTM network to obtain the newly generated memory matrix, where the  $i$ -th vector expression is shown in Eq. (10)[21].

$$H_k^{t+1} = \begin{cases} RNN(f_{MLP}(h_k^t)) \\ RNN(f_{neighbor}(h_i^t, h_k^t)) \end{cases} \quad (10)$$

Where  $RNN$  represents the fully connected recurrent neural network,  $h_k^t$  represents the vector values at the moment, and  $f_{MLP}$  is the LSTM network. The process focuses on the nearest neighbor function of vector  $j$  and vector  $k$ , and this attention algorithm is formulated in Eq. (11).

$$f_{neighbor}(h_i^t, h_k^t) = \frac{1}{K} \sum_{k \in K} a_{ij}^k f_k(h_i^t, h_k^t) \quad (11)$$

Where  $a_{i,j}^k$  represents the threshold value for simplifying the graph structure. The study uses a deep knowledge tracing model to obtain vector graphs, which are acyclic structural graphs with positive-valued full connectivity between the nodes of the graph, and a threshold parameter will be introduced to simplify the graph structure. When the topic concept at the moment  $t$  is  $K$ , only the vector  $i$  adjacent to it needs to be considered, and the graph model is defined using the connection relations between directed graphs. Finally, the model needs to encode the knowledge concepts. The knowledge concept sailing in at  $t+1$  is set to  $e_{t+1}$ , and forward neural network processing is performed to obtain  $S_{t+1}$ . This encoding long queue is  $d$ , which is calculated by equation (12).

$$S_{t+1} = \sigma(W_1 \cdot \sigma(W_o \cdot e_{t+1} + b_o) + b_1) \quad (12)$$

Where the forgetting gate in the structure of the feedforward neural network is represented with the previous memory gate input parameter [22]. At the moment of  $t+1$ , the memory matrix vector  $H$  is multiplied by the vector points of the knowledge encoding  $S$  to obtain the true mastery probability value for moment  $P_{t+1}$ . The study developed a methodological model of reinforcement learning by mimicking the teacher's learning behavior towards the students and using a deep knowledge tracking model process sample training. In the process of using the knowledge tracking model, the recommended questions are first selected and then the input of the questions is completed. The predicted answers are obtained and the concept of knowledge is encoded, and this encoding is used as the input learning state vector value for training the reinforcement learning method. The model structure is shown in Fig. 4.

As shown in Fig. 4, when the algorithm is executed, a random state vector of fixed length is first initialized and the basic level of knowledge point concept mastered by the student is identified, where each element of the vector is labeled as the degree of learning to the knowledge value. On this basis, a convolutional neural network is built. Then, the process of learning action selection is executed, where three outputs exist in this process. When the return value is 1, then the learning is sufficient and no subsequent selection of knowledge-related topics is required. When the return value is 0, the old and new states, actions and their corresponding reward values need to be combined and saved to the experience pool according to the fixed collocation principle as the basis for subsequent learning and optimization. Continuous action selection for the immediate moment is performed, and the deep knowledge tracking model is trained for deep reinforcement learning after several iterations of the loop. The reinforcement learning method designed for this research uses the Nature DQN (Deep Q Network, DQN) network structure [23]. Two neural networks with the same number of layers and nodes were set up to address the correlation between the data samples and the

network before training. One of the  $Q$  networks is used as the training network with the input value of the state value of the environment mention training and the output corresponding to it is the recommended action. The second  $Q$  network is used as the target network to reduce the overfitting of the neural network, so that this network serves as the final target value for reinforcement learning, thus enabling an optimal update of the corresponding network parameters. The study evaluates the model in terms of average accuracy, average completeness and accuracy, where the average completeness ( $AP$ ) is calculated by equation (13).

$$AP = \sum_{u=1}^r P(u) \quad (13)$$

Where  $r$  denotes the number of categories and  $u$  denotes the category labels. The average finding rate ( $AR$ ) is expressed by Eq. (14).

$$AR = \sum_{u=1}^r R(u) \quad (14)$$

The accuracy of ( $ACC$ ) is calculated by Eq. (15) [24].

$$ACC = \frac{(TP + TN)}{(TP + FN + FP + TN)} \quad (15)$$

Where  $TP, FP$  denote the number of positive classes and the number of negative classes for positive class prediction, respectively.  $FN, TN$  denote the number of positive classes and the number of negative classes for negative class prediction, respectively. In addition, the study determines the performance of the model by comparing the Area Under Curve (AUC) of each dataset completion curve.

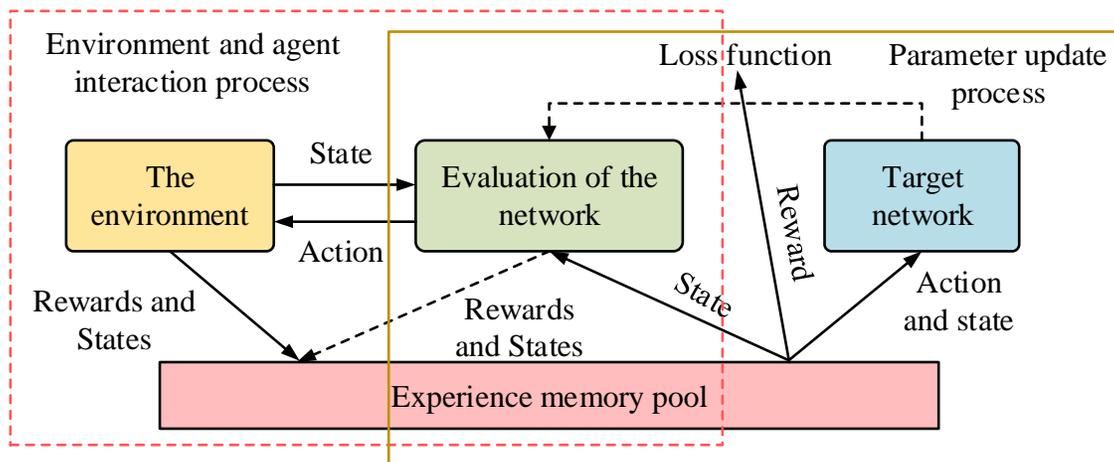


Fig. 4. Structure diagram of training reinforcement learning method model.

### III. ANALYSIS OF EXPERIMENTAL RESULTS OF A DEEP DIRECTED GRAPH KNOWLEDGE-TRACKING-BASED MODEL FOR DISTANCED PERSONALIZED ENGLISH TEACHING

#### A. Analysis of the Effectiveness of a Deep Learning Knowledge Tracking Model based on Multidimensional Feature Fusion

The validity, interpretability analysis of the deep knowledge tracking model incorporating multidimensional features was conducted. In this experiment, the Adam optimizer was used to train the model, and all datasets were sliced 7:3 to test set and training set with 10 training times. Based on tensor flow and keras, the methods in Table I were completed in NVIDIA 1080 Ti GPU environment. At present, knowledge tracing methods widely used include BKT (Bayesian knowledge tracing), DKT (Deep knowledge tracing), and DKVMN (Dynamic key tracing) value memory networks) these three knowledge tracking methods. BKT can construct the knowledge state of learners as a set of binary variables, but its prediction accuracy depends on the experience of the teacher, and the degree of automation is low. DKT is the

current mainstream knowledge tracking method without a lot of teaching experience and manual labeling, but it is difficult to obtain more specific knowledge state of learners. DKVMN can solve the problems existing in BKT and DKT with high prediction accuracy, but it ignores the factors and characteristics existing in the learning process of learners. AUC (Area Under Curve) ranges from 0.5 to 1.0. The closer the AUC is to 1.0, the higher the authenticity of the detection method. Table I shows the AUC values of multiple algorithms on 6 datasets.

As can be seen from Table I, the DKVMN-BORUTA model proposed in the study has an AUC value of around 0.83, which is higher than that of other three models, indicating the superior performance of the studied model. Since the model incorporates multidimensional features in the learning process, it can to some extent solve the problem of traditional knowledge tracking models for modeling simple mathematical logic in learning and achieve better simulation of the real state of the learner as well as better assessment of the learner's knowledge state.

TABLE I. AUC VALUES OF FOUR METHODS ON DATASETS

Data set	Number of students	Knowledge tag	AUC			
			BKT	DKT	DKTVMN	DKVMN-BORUTA
ASSIST2014	4208	112	0.632	0.806±0.02	0.815±0.01	0.832±0.01
ASSIST2015	19850	110	0.646	0.726±0.01	0.748±0.01	0.804±0.02
Statics2011	334	158	0.71	0.803±0.02	0.799±0.02	0.832±0.01
Synthetic	2000	6	0.761	0.804±0.01	0.806±0.01	0.841±0.02
Khan Academy	17825	108	0.689	0.802±0.02	0.814±0.02	0.822±0.01
KDD Cup	3996	180	0.743	0.783±0.01	0.811±0.01	0.831±0.02

The study collected data from the first half semester of a university’s personalized English course instruction in 2021 and organized that data into a dataset, which was used to train the DKVMN-BORUTA model. The questions were set with good differentiation and moderate difficulty, and if the mastery rate of the knowledge point under this question set was less than 40%, it indicates that the learners did not master the knowledge point. To verify the reliability of the proposed method, a random sample of learners was selected for the topic mastery statistics, and the changes in the knowledge state of the learners are shown in Fig. 5(a). Also, to verify whether the proposed method can achieve the goal of personalized instruction, the randomly selected learners were divided into three groups. No knowledge state information was provided for group A, knowledge state information but no precise instructional support service was provided for group B, and knowledge state information with precise instructional support service was provided for group C. Then, the learning performance and learning review time of the learners in the three groups were recorded, as shown in Fig. 5(b).

As can be seen from Fig. 5(a), the overall trend of learners’ mastery of the knowledge points was on the rise, indicating that this learner’s learning status became gradually become after being taught by the proposed method, indicating that the proposed knowledge tracking method has better readability. As shown in Fig. 5(b), the average test scores of Group C and Group B were higher than that of Group A, and the average review time of both groups was less than that of Group A. This

indicates that the deep learning knowledge tracking model based on multidimensional feature fusion proposed by the study is effective and can find learners’ weak points for learning knowledge more precisely. The average score of group C is 2.5 points higher than group B, and the average revision time is 4h less than group B. This indicates that the deep knowledge tracking model has better interpretability and can fully combine the multidimensional features of learners’ learning process to reach accuracy teaching and improve the English learning efficiency.

*B. Analysis of the Effectiveness of a Distanced Personalized English Teaching Model F incorporating Deep Directed Graph Knowledge Tracing and Graph Structure Rules*

After verifying the effectiveness of the deep learning knowledge tracking model based on multidimensional feature fusion, an experimental analysis of this deep learning knowledge tracking model combined with graph structure rules for distanced personalized English teaching and learning was conducted. The same dataset as before was used in this experimental phase and its basic situation is shown in Table II.

As can be seen from Table II, the lowest AUC value of the proposed model is 0.893 and the highest is 0.921, which shows the effectiveness and superior performance of the proposed model. Then, the performance of the proposed model was compared with other three knowledge tracking models in terms of prediction effectiveness by measuring the AUC of each dataset and its variation with the number of training sessions, and the results of the comparison are shown in Fig. 6(a).

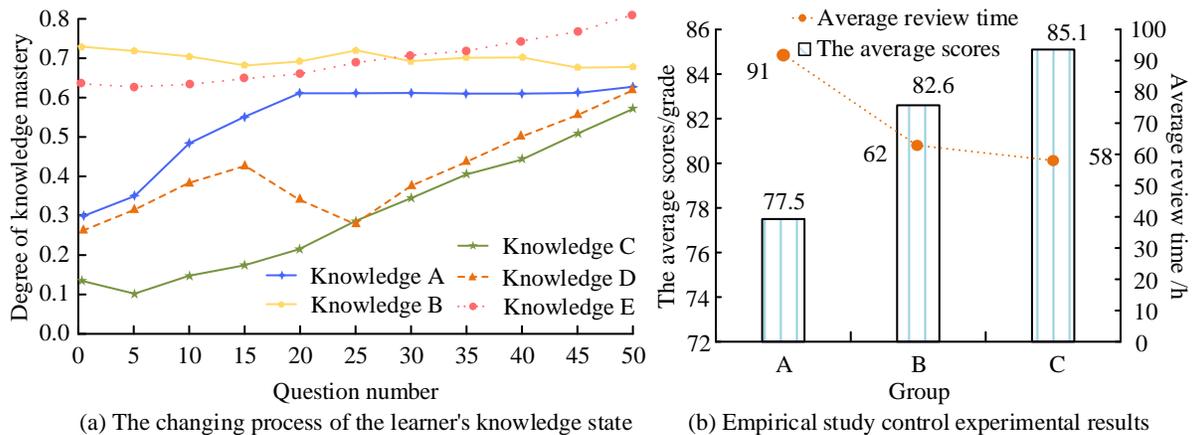


Fig. 5. The changes of learners’ knowledge state and the results of the empirical study.

TABLE II. AUC VALUES OF SIX PUBLIC DATA SETS AND FOUR MODELS

Data set	Number of students	Knowledge tag	Number of interactions	Number of valid result records	The AUC of the study model
ASSIST2014	4296	120	325682	314126	0.914±0.1
ASSIST2015	18246	116	598521	565247	0.893±0.2
Statics2011	400	188	3548	3264	0.912±0.1
Synthetic	2100	10	14621	13346	0.911±0.1
Khan Academy	18364	127	236489	216721	0.921±0.1
KDD Cup	4010	186	11000	10834	0.898±0.2

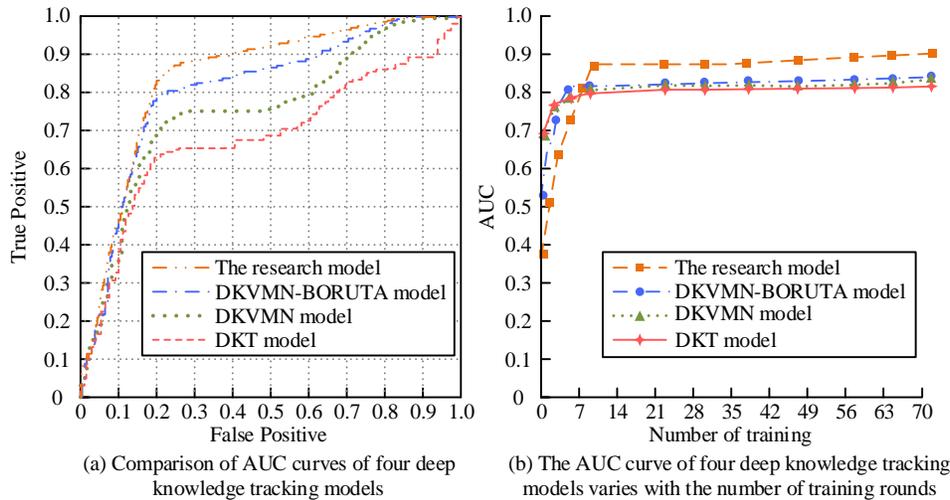


Fig. 6. AUC curves of four deep knowledge learning models.

From Fig. 6(a), it can be seen that the model proposed in the study had an average AUC value of 0.92 in various datasets, while the AUC value of the DKT model was only 0.73. Through comparing the boundary baseline on the KDD dataset, the model proposed in this study improved the gain by a factor of about four. As can be seen from Fig. 6(b), it can be found that the proposed model tended to be stable after the 9th training session, and the AUC value of the proposed model was consistently higher than that of the other three models after the 9th training session, indicating the superiority of the proposed model. The study collected the learners' mastery degree of each knowledge point in the personalized English instruction course, and the collected data were processed to obtain the empirical findings, as shown in Table III.

As shown in Table III, the prediction accuracy of the proposed method was 94.3%, which was 4.9%-26.0% higher than that of the other models. The mean square error of the proposed model was 0.1623, which was the smallest among the tested models. In addition, the F1 score of the proposed model was 0.92, which was the highest among the tested models. The

research data indicated that the proposed method had strong performance. Finally, the study tested the variation of the reward value and the number of training sessions, as shown in Fig. 7(a). The comparison of the overall English scores of the learners in each group and the time spent on revision is shown in Fig. 7(b).

From Fig. 7(a), it can be seen that the overall accuracy of the intelligent personalized recommendation algorithm based on reinforcement learning proposed in this study tended to increase with the increase of training times, and the difference between it and the random recommendation algorithm was gradually increasing, with the difference in the range of 6%-9%. This result indicates that the proposed method had more accurate recommendation capability, which can meet the demand of distanced personalized English teaching. It can be seen from Fig. 7(b) that the comprehensive score of the study group using the study method was 97.3, and the average daily review time of the students in this group was 63min. Research shows the proposed method can effectively improve learners' overall English performance and save learning time.

TABLE III. EMPIRICAL RESEARCH RESULTS

Model type	AP (%)	ACC (%)	F1-score	MSE
BKT	87.3	68.3	0.73	0.2491
DKT	86.2	77.9	0.79	0.1892
DKVMN	88.4	82.6	0.82	0.1834
DKVMN-BORUTA	88.6	89.4	0.89	0.1729
The research model	90.7	94.3	0.92	0.1623

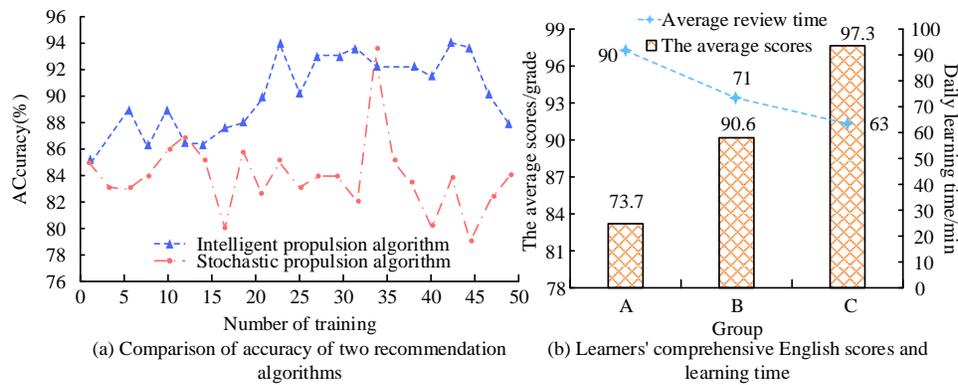


Fig. 7. Comparison of recommendation algorithm accuracy with learners' English scores.

#### IV. CONCLUSION

The rapid development of Internet information technology has led to the continuous innovation and progress of English online education models. This study proposes a distanced personalized English teaching model incorporating deep directed graph knowledge tracking. Firstly, with the help of BORUTA feature selection model, the multidimensional features of the learning process are mined and a deep knowledge-tracking model of the multidimensional features of the learning process is fused. This optimized deep knowledge tracking method is then combined with graph structure rules to construct a fused approach-based model for distanced personalized English teaching and learning. The results show that the proposed model has the lowest AUC value of 0.893 and the highest of 0.921 on each dataset, indicating its effectiveness and superior performance. The prediction accuracy of the proposed model is 94.3% and its mean square error is 0.1623, which is the smallest among the models. The F1 score of the proposed model is 0.92, which is the highest among the models indicating that the model has strong performance. Moreover, the overall accuracy of the proposed method shows an increasing trend and is higher than that of the random recommendation algorithm. The optimized deep knowledge tracking method based on the multi-dimensional features of the learning process has practical teaching application value. The fusion model has higher recommendation accuracy and better recommendation effect than the traditional method. And it can meet the needs of personalized remote English teaching and targeted recommendation teaching, so as to improve the quality of English teaching. This study provides a practical theoretical basis and reference direction for online teaching methods in the future. However, there is still room for improvement in the interpretability of the model, and future study will focus on the improvement and enhancement of the performance of the model.

#### REFERENCES

[1] M. D. Labitad, L. S. Lomibao, The Experiences, Challenges and Perception on Online Lesson Study (O-LS) as Teacher Professional Development Program. *American Journal of Educational Research*, vol. 9(10), pp. 639-646, 2021.

[2] A. Cully, Y. Demiris, "Online Knowledge Level Tracking with Data-Driven Student Models and Collaborative Filtering". *Institute of Electrical and Electronics Engineers (IEEE)*, vol. 32(10), pp. 2000-2013, 2020.

[3] Wang M M. Educational Probation Under the Tutorial System and the Construction of English Student Teachers' Knowledge. *teaching English in America and China: English version*, vol. 16(2), pp. 75-83, 2019.

[4] T. Y. Zhao, M. Zeng, J. H. Feng, "An Exercise Collection Auto-Assembling Framework with Knowledge Tracing and Reinforcement Learning", *Journal of Computer Science and Technology*, 2022, vol. 37(5), pp. 1105-1117, 2022.

[5] T. Dani, R. Priyatikanto, A. Winarko, P. Gerhana, "Solar Wind Speed Time-Series Forecasting Based on Long Short-Term Memory (LSTM) Neural Network Model". vol. 275, pp. 193-198, 2022.

[6] Gervett, Koedingerk, Schneider J. When is deep learning the best approach to knowledge tracing. *journal of Educational Data Mining*, vol. 12(3), pp. 31- 54, 2020.

[7] Sapountzia, Bhulais, Corneliszi. "Dynamic knowledge tracing models for large-scale adaptive learning environments" *International Journal on Advances in Intelligent Systems*, vol. 12, pp. 93-110, 2019.

[8] S. U. Hassan, H. Waheed, N. R. Aljohani, M. Ali, S. Ventura, F. Herrera, "Virtual learning environment to predict withdrawal by leveraging deep learning", *International Journal of Intelligent Systems*. vol. 34(8), pp. 1935-1952, 2019.

[9] A. A. Mubarak, H. Cao, S. A. Ahmed, "Predictive learning analytics using deep learning model in MOOCs' courses videos", *Education and Information Technologies*. vol. 26(1), pp. 371-392, 2021.

[10] Käsert, Klinglers, Schwingag. "Dynamic Bayesian networks for student modeling". *IEEE Transactions on Learning Technologies*, vol. 10(4), pp. 450-462, 2019.

[11] Deonovicb, Yudelso nm, Bolsinovam. "Learning meets assessment. *behaviormetrika*," vol. 45(2), pp. 457-474, 2018.

[12] H. Yang, L. P. Cheung, "Implicit heterogeneous features embedding in deep knowledge tracing." *cognitive Computation*. vol. 10(1), pp. 3-14, 2018.

[13] Z. Wang, X. Feng, J. Tang, "Deep knowledge tracing with side information", *Proceedings of the International Conference on Artificial Intelligence in Education*. cham: springer, vol. 2, pp. 303-308, 2019.

[14] J. J. Vie, H. Kashima, "Knowledge tracing machines: factorization machines for knowledge tracing. in *Proceedings of the AAAI Conference on Artificial Intelligence*", vol. 33(1), pp. 750-757, 2019.

[15] X. Tian, F. Liu, "Capacity tracing-enhanced course recommendation in MOOCs", *IEEE Transactions on Learning Technologies*. vol. 14(3), pp. 313-321, 2021.

[16] Y. Zou, X. Yan, W. Li, "Knowledge Tracking Model Based on Learning Process", *Computers and Communications*, vol. 10, pp. 7-17, 2020.

[17] Z. Zhang, X. Liu, Z. Li, H. Hu, "Outburst prediction and influencing factors analysis based on Boruta-Apriori and BO-SVM algorithms", *Journal of Intelligent & Fuzzy Systems: Applications in Engineering and Technology*, vol. 41(2), pp. 3201-3218, 2021.

[18] A. Es, B. Rw, B. Aks, "Securing color image transmission using compression-encryption model with dynamic key generator and efficient symmetric key distribution", *Digital Communications and Networks*, vol. 6(4), pp. 486-503, 2020.

- [19] F. M. Faulstich, R. Kim, Z. H. Cui, "Pure State  $v$ -Representability of Density Matrix Embedding Theory", *Journal of chemical theory and computation: JCTC*, vol. 18(2), pp. 851-864, 2022.
- [20] J. Zhang, Q. Xu, "Attention-Aware Heterogeneous Graph Neural Network", *Big Data Mining and Analytics*, vol. 4(4), pp. 233-241, 2021.
- [21] X. Shi, Y. Li, Y. Yang, B. Sun, F. Qi, "Multi-models and dual-sampling periods quality prediction with time-dimensional K-means and state transition-LSTM network", *Information Sciences: An International Journal*, vol. 580, pp. 913-933, 2021.
- [22] M. Zheng, J. Luo, Z. Dang, "Feedforward Neural Network Based Time-Varying State-Transition-Matrix of Tschauner-Hempel Equations", vol. 69(2), pp. 1000-1011, 2021.
- [23] Y. S. Lee, A. Masood, W. J. Noh, S. G. "Cho DQN based user association control in hierarchical mobile edge computing systems for mobile IoT services". *Future generation computer systems*, vol. 137, pp. 53-69, 2022.
- [24] I. Skrzypczak, G. Oleniacz, A. Leśniak, K. Zima, M. Mrówczyńska, K. Kazak Jan "Scan-to-BIM method in construction: Assessment of the 3D buildings model accuracy in terms inventory measurements", *Building Research & Information*, 2022, vol. 50(8), pp. 859-880, 2022.

# A Visual Target Representation using Saliency Detection Approach

Shekun Tong\*<sup>1</sup>, Chunmeng Lu<sup>2</sup>

College of Information Engineering, Jiaozuo University,  
Jiaozuo, Henan 454100, P. R. China<sup>1</sup>  
College of Artificial Intelligence, Jiaozuo University,  
Jiaozuo, Henan, 454100, P. R. China<sup>2</sup>

**Abstract**—The task of saliency detection is to identify the most important and informative part of a scene. Saliency detection is broadly applied to numerous vision problems, including image segmentation, object recognition, image compression, content-based image retrieval, and moving object detection. Existing saliency detection methods suffer a low accuracy rate because of missing components of saliency regions. This study proposes a visual saliency detection method for the target representation to represent targets more accurately. The proposed method consists of five modules. In the first module, the salient region is extracted through manifold ranking on a graph, which incorporates local grouping cues and boundary priors. Secondly, using a region of interest (ROI) algorithm and the subtraction of the salient region from the original image, other parts of the image, either related or nonrelated to the interested target, are segmented. Lastly, those related and non-related regions are classified and distinguished using our proposed algorithm. Experimental result shows that proposed salient region accurately represent the interested target which can be used for object detection and tracking applications.

**Keywords**—Saliency detection; target representation; vision system; object detection

## I. INTRODUCTION

Saliency detection methods can be categorized into fixation prediction-based and salient object detection methods. Fixation prediction-based methods are strongly related to biological models. Their processes simulate how human fix locate an image. On the other hand, object-oriented approaches generate maps such that salient regions uniformly cover the whole objects [1].

Fixation prediction methods are inspired by biological principles. Itti et al. [2] proposed the seminal bottom-up saliency model derived from the human visual selective attention mechanism. It obtains saliency maps in different feature channels with a center-around operation and combines them linearly. Hou and Zhang [3] presented a saliency model counting for the Fourier envelope and the differential spectral components, called the spectrum residual, to extract salient regions. Garcia-Diaz et al. [4, 5] propose an adaptive approach achieved by decorrelation and contrast normalization. These fixation prediction-based methods usually overemphasize local contrast and difference from the neighborhood. They are more likely to produce spotlight saliency maps with low resolution and high saliency values on the object boundaries.

On the other hand, Achanta et al. [6] compute the saliency likelihood of each pixel based on its color contrast to the entire image. Ming et al. [7] consider the global region contrast concerning image and spatial relationships across the regions to extract a saliency map. In [8], Goferman et al. simultaneously model local low-level clues, global considerations, visual organization rules, and high-level features to highlight salient objects with their contexts. Such methods using local contrast tend to produce higher saliency values near edges instead of uniformly highlighting salient objects.

This paper proposes a new method for saliency detection for target appearance representation in images. It incorporates visual features and spatial information with the guidance of prior saliency knowledge. To provide more accurate visual cues, region descriptors are introduced for image segments by computing two saliency measures, feature distinctiveness, and spatial distribution. In contrast to previous models, which linearly combine basic features for visual cues, we provide nonlinear integration of the features. In addition, by taking the advantage of the prior saliency distribution obtained from a convex hull of salient points, we heighten the contrast of foreground and background [1].

In [9], a bottom-up method is proposed to detect salient regions in images through manifold ranking on a graph, which incorporates local grouping cues and boundary priors. This method adopts a two-stage approach with the background and foreground queries for ranking to generate the saliency maps. This method is taken advantage of node construction for saliency map identification. Fig. 1 shows the node construction in the GMR method.

Although various saliency detection methods are proposed and present promising results, current methods are not feasible and effective when an object is composed of several parts [1]. For example, as shown in Fig. 2, our model fails to detect the wheels since they are discriminated from the main part of the car. Therefore, current saliency detection methods are not effectively able to represent an object, which makes fail object appearance representation in moving object detection systems. Therefore, it is required to investigate an effective approach to deal with this issue. Fig. 3 shows the saliency fail detection in current methods.

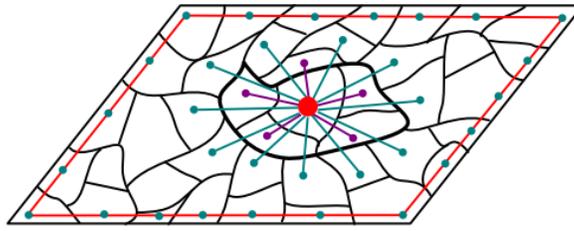


Fig. 1. Nodes construction in GMR [9]

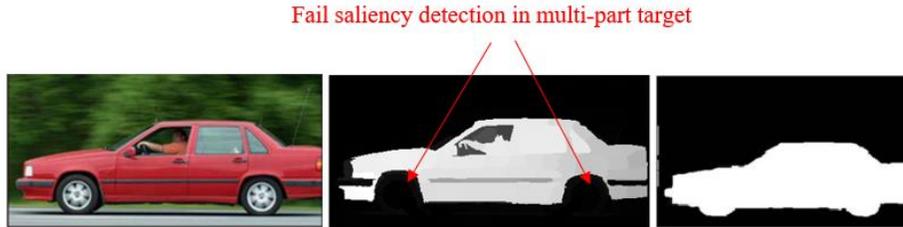


Fig. 2. Fail saliency detection in GMR multi-part saliency detection method. The first image is the original image, the second is the result of saliency map for [9] method and the last image is grand-truth

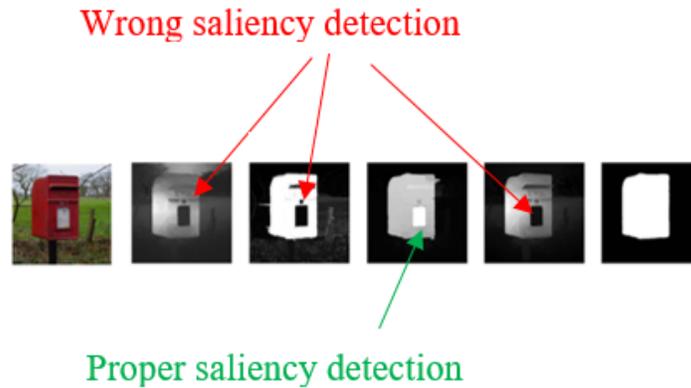


Fig. 3. Wrong and propose saliency detection in different saliency detection methods. The first image is original image, the last image is grand-truth images, and the others from left to right are the saliency map results for LDRCCB [10], SF [11], BS [12], GMR [9]

## II. PROPOSED METHOD

In dealing with the mentioned problem in the previous step, a saliency detection method is proposed to recognize saliency components for a salient region. This component of saliency recognition aims to represent the accurate region because sometimes they consist of several parts, as shown in Fig. 4, and they are required to recognize. The proposed method is structured by five modules, salient region detection, salient region localization, region of interest extraction, particle region segmentation, and belonging ratio (BRM) measurement. The proposed method for multi-part saliency recognition is shown in Fig. 4. The detail of each step is discussed in the following sub-sections.

### A. Salient region Detection

There are numerous region saliency detection methods. This study adopted a graph-based manifold ranking (GMR) method for salient region detection [17]. As reported in [13], the GMR methods presented promising results for salient region detection [9]. The source code of the GMR method is

available from the author's website<sup>1</sup>. The source code collects a saliency map of the GMR method. For example, Fig. 5 shows the saliency map (saliency result) using the GMR method. The original image (Fig. 5(a)) was taken from the MSTR-1000 dataset, which is available from the author's website.

### B. Salient region Localization

After salient region detection, the localization of the salient region is performed using centroid finding and region size identification [18]. The purpose of localization is to locate the region saliency on the original image. This localization is also used to generate an expanded region to cover different parts of an interested target.

1) *Saliency map binarization:* To do the localization, the saliency map image is firstly converted to zero 0 and 1 pixels (binarization) using thresholding. The Otsu operator uses for thresholding and binarization. Fig. 6 shows the binarization of the saliency map image.

<sup>1</sup> [http://ivrlwww.epfl.ch/supplementary\\_material/RK\\_CVPR09](http://ivrlwww.epfl.ch/supplementary_material/RK_CVPR09)

2) *Centroid identification*: In this step, the centroid of the binary saliency map is identified. The connected component function is used to extract white pixels as a blob. Blob properties are used to recognize the center of the shape. A geometric decomposition algorithm [14, 15] is used to find the centroid identification. In geometric decomposition algorithm, the centroid of a blob figure as  $X$  can be computed by decomposing it into certain number of parts as,  $X_1, X_1, X_1, \dots, X_n$ , therefore, the centroid and area  $A_i$  of each part, and then the coordination of  $C_x$  and  $C_y$  can be computed as equations (1),

$$C_x = \frac{\sum C_{ix}A_i}{\sum A_i}, C_y = \frac{\sum C_{iy}A_i}{\sum A_i} \quad (1)$$

Using above equation, the centroid of salient region can be obtained. The centroid is represented by a coordination that can be used for further process.

3) *Region size identification*: Bed on the output of the previous step, which was the identified centroid, we can access the centroid coordination of the salient region. Other geometry properties such as the height and width parameters can be specified by this coordination. In this regard, considering  $S_{cp}(C_x, C_y)$ , where  $S_{cp}$  is the center point of salient region,  $C_x, C_y$  are the values for  $X$  and  $Y$  axis. Using the center point and identified height and width, an expanded region with  $H'$  and  $W'$  are calculated by following equation,

$$W' = W + 1/n (W) \quad (2)$$

$$H' = H + 1/n (H)$$

where the  $W'$  and  $H'$  are the expanded values by the  $n$  variable,  $n$  is a variable for expanding factor which is set 5 in this experimental. The expanding factor was obtained based on different experiments, assume as  $n = 6$ . Fig. 7 shows expanded height and weight of salient region with corresponding center point,  $W, H, W'$  and  $H'$ .

4) *Region of interest extraction*: The output from the previous step includes information such as expanded region details and salient region center points. This information is used to extract and generate a region of interest (ROI) from a binary saliency map. The extracted region of interest of the binary saliency map is defined as the ROI of the binary saliency map. Moreover, according to ROI, a binary saliency map can be extracted from the saliency map and original image using mapping of the extracted binary saliency map. The extracted new regions from the saliency map and original image are defined as the ROI of the saliency map and the ROI of the original image, respectively. Fig. 8 shows the region of interest extraction for the original, saliency map, and binary saliency map images.

5) *Particle region segmentation*: In this step, the ROI is segmented into different regions called particles. The purpose of particle region extraction is to extract different regions, including regular and symmetric shapes, from the ROI of the original image. The extracted particles are then transferred to the next step to check whether the particle belonged to the target (the detected salient region) or not. To segment different

particles, the following steps are considered; edge segmentation, image enhancement, filling blobs, and applying the proposed shape descriptor. Fig. 9 shows the steps of particle region segmentation.

### C. Thresholding and Edge Segmentation

Edges are significant properties of each region. The edges are used to detect region boundaries and segment the region. The segmented region using edges, can be used in more processes, such as region analysis and recognition. In this study, the edges are also used to segment different regions on the ROI of the original image and then generate particles. The particles are defined as different regions that are generated after edge segmentation.

1) *Region enhancement and filling blobs*: The detected edges from previous steps are contained many disconnected regions. Image enhancement is required to close the regions and remove some noises [19]. Therefore, noise removal [20], mathematical morphology operator [21], lightweight operators including dilation and closing [22], are applied to enhance the result of the edge detection process. Then filling function is applied, to fill the connected blobs in the enhanced image. Fig. 9(c) shows the image result from the image enchantment step.

2) *Interested region shape extraction*: As shown in Fig. 9(c), the result of image enhancement is still contained many noises, which are required to remove. The noises are involved different types of shapes. To remove these noises, the proposed shape descriptor is applied in this image. Additionally, regular [23] and symmetry [24] regions called interesting shapes (such as wheels) are extracted based on the proposed shape descriptor. With the proposed shape descriptor, the particle regions with regular and symmetry shapes are detected. Since, the final target (the salient region) that we are looking for (such as vehicles) are involved regular components. Fig. 9(d) shows the particle regions result after the interested shape extraction process.

3) *Saliency components recognition*: As discussed in the previous step, the result of the interest shape extraction step is contained regular and symmetry region shapes, which can be belonged to a salient. The belonged regions define as saliency components. To recognize the saliency components, an algorithm is proposed, which is shown in Table I. In the proposed algorithm, two input images are required. The first input image is a set of particles, including the segmented particles shown in Fig. 10. The other input is a saliency mask. The interior regions of a binary saliency map are filled with this saliency mask. Fig. 11 shows the saliency mask image. The saliency mask is considered because the exterior parts of the salient region are required to be recognized.

The details of proposed algorithm describe as follows.

### D. Particle Properties Identification

In this step, the properties of each particle are identified. To do this identification, the edges for particle regions and salient mask are detected, and their region is extracted. The Sobel

edge detector is used for edge segmentation. Fig. 12 shows the edge segmentation of the salient mask. The segmented regions are then integrated into one image. Fig. 13 shows a man-made image, including the integration of a salient mask and some particles regions.

To identify the properties of the particles, the centroid of particle is firstly extracted by taking the particle area and dividing it up into differential areas. The arbitrary shape has an area denoted by  $A$ . Differential area  $dA$  that exists some distance  $x$  and  $y$  from the origin. The total area is denoted as  $\int dA$ , which is the first moment of area in each direction from the following Eq. [16]:

$$Q_x = \int_A y dA \quad (3)$$

$$Q_y = \int_A x dA$$

The first moment of area is the integral of a length over an area. It is important because it helps us to locate the centroid of the particle. According to the obtained area, the centroid is defined as the "average  $x$  (or  $y$ ) position of the area"

$$\bar{x} = \frac{Q_y}{A} = \frac{\sum_i \bar{x}_i \bar{A}_i}{\sum_i \bar{A}_i} \quad (4)$$

$$\bar{y} = \frac{Q_x}{A} = \frac{\sum_i \bar{y}_i \bar{A}_i}{\sum_i \bar{A}_i}$$

Fig. 5 to 15 shows the centroid of the particles using calculation of centroid moments. Using the identified center point, diameter for particle is identified, as shown in Fig. 14.

#### E. Particle Mask Generation

In this step, a mask is generated according to the shape of the particle inspired from [25]. The purpose of this mask is to determine the range for exploring the area. The exploring area (ExA) is a search range to check whether the particle belongs to the target (salient region) or not. The diameter of the mask depends on particle diameter. It calculates by  $D_{mask} = 2 * D_{particle}$ . Fig. 15 shows the generated particle masks. Furthermore, Fig. 16 shows ExA on integrated salient and particles image.

#### F. Confluence Lines (CL) Generation

Confluence lines (CL) are the lines to find intersection points in different regions. These regions are the particles and the salient region. The CL lines are started from the particle center point and end up on the mask. Therefore, it is required to identify the endpoints on the mask to draw these lines. Having the area value of the mask to find the endpoint on the mask or particle is required. The area of the particle mask is considered that can be calculated using Eq. (3).

$$A = \int_0^\theta \int_0^r dS = \int_0^\theta \int_0^r \tilde{r} d\tilde{r} d\tilde{\theta} = \int_0^\theta \frac{1}{2} r^2 d\tilde{\theta} = \frac{r^2 \theta}{2} \quad (5)$$

where  $\theta$  is denoted as an angle between two individual points on the mask region, the radius of the mask,  $S$  is the region mask, and  $A$  is the area of the mask. Using the identified area, a mask can be divided into equal sectors. In this regard, several sectors and angles between sectors, are required. The equation is used to address the angle between each sector.

$$\varphi = \frac{2\pi}{n} \quad (6)$$

where  $n$  is number of sectors, and  $\theta$  denotes the angle between CL lines (sectors). According to our experiment, with 20 sectors, we can get promising results. Therefore, the mask contains 20 points (since there are 20 equal sectors on a mask). These points are considered endpoints for CL lines in which the angle between the CL lines is as,  $\theta = \frac{\pi}{10}$ . The angle between each pair CL is 18. Fig.17 shows a circle division demonstration inspired by trigonometry constants on real radicals<sup>2</sup>. Fig. 18 illustrates the sectors and confluence lines (CL) in a mask.

#### G. Particle and Salient Region Distance Measurement

Using the CL lines, the intersection points from the salient region and particles with CL can be extracted. The points form a salient region, and particles are collected; and composed of a pair of two points used for distance measurement. Fig. 19 shows the point on the salient region as  $P_s$  and particles as  $P_p$ . Euclidean distance (Equation (5)) is used for distance measurement. Then, Fig. 20 shows the result of Euclidean distance for distance measurement between  $P_s$  and  $P_p$ .

$$D_i = \sqrt{\sum_{i=1}^n (P_s - P_p)^2} \quad (7)$$

where, the  $P_p$  and  $P_s$  are defined as,

$P_p (x_p, y_p) \rightarrow$  intersection point for CL and particle, and

$P_s (x_s, y_s) \rightarrow$  intersection point for CL and salient region

The obtained distances for each pair of  $P_s$  and  $P_p$  called  $D_i$  are compared. The purpose of this comparison is to identify a number of equal lines and non-lines distances. For identification number of equal and equal distances, we consider conditions as shown in Eq. (6),

$$\text{For each } D_i \text{ and } (D_i - 1), \text{ if } \begin{cases} D_i - (D_{ps} - 1) > \varepsilon \rightarrow N_{eq} = N_{eq} + 1 \\ D_i - (D_{ps} - 1) < \varepsilon \rightarrow N_{non-eq} = N_{non-eq} + 1 \end{cases} \quad (8)$$

#### H. Possession Ratio Measurement

In this step, a ratio is introduced to decide whether the particle is belonged to the silent region (target) or not. The ratio is called the possession ratio (or belonging ratio) ( $\mu_{BR}$ ). Eq. (7) is defined in this study to measure the  $\mu_{BR}$  for each particle. A condition rule then is defined to decide the particle is belonged, which is based on a threshold denoted as  $\delta$ .

$$\mu_{BR} = 1/n(\sum_{i=1}^n N_{eq}) \quad (9)$$

where  $N_{eq}$  is a counter variable to count the equal distances and  $n$  number of lines. Based on obtained value from Eq. (7), a condition defines as shown in Eq. (10), to decide whether the particle is belonged to the target (salient region) or not as,

$$\begin{cases} \mu_{BR} = 1 & \text{then } P_i \leftarrow T \\ \mu_{BR} > \delta \text{ and } \mu < \delta + 0.1 & \text{then } P_i \leftarrow T \\ \mu_{BR} < \delta & \text{then } P_i \leftarrow N \end{cases} \quad (10)$$

According to calculated  $\mu_{BR}$  and asses the condition, the particle is labeled as True ( $T$ ) or Negative ( $N$ ). Lastly,

<sup>2</sup>[https://commons.wikimedia.org/wiki/File:Unit\\_circle\\_angles\\_color.svg](https://commons.wikimedia.org/wiki/File:Unit_circle_angles_color.svg)

according to  $T$  and  $N$  labels, the corresponding particles are recognized as the saliency components.

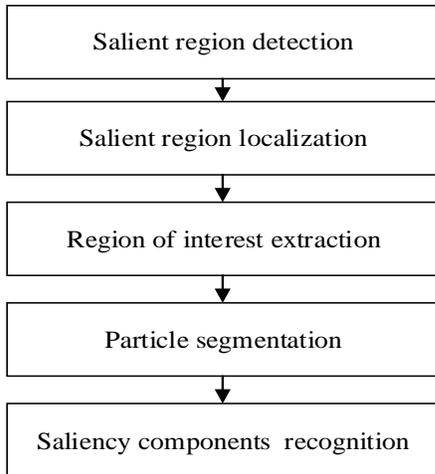
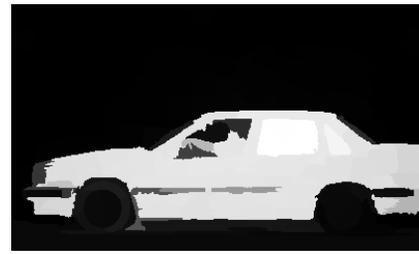


Fig. 4. The proposed method for multi-parts saliency recognition



(a)



(b)

Fig. 5. The saliency map from GMR method, (a): original image, (b): saliency map

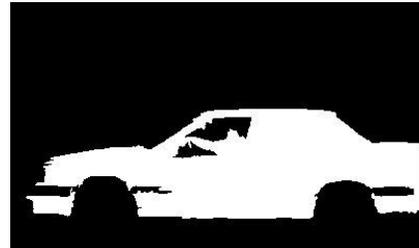


Fig. 6. Binarization of saliency map

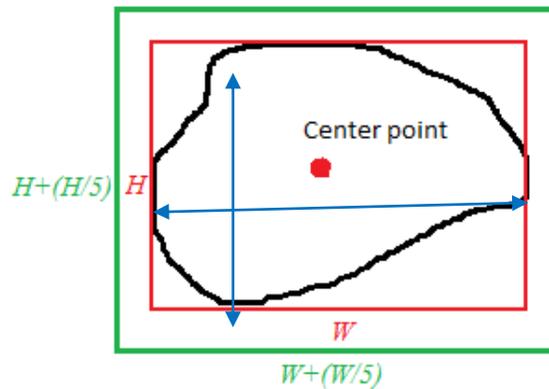


Fig. 7. Expanded height and weight of salient region with corresponding center point

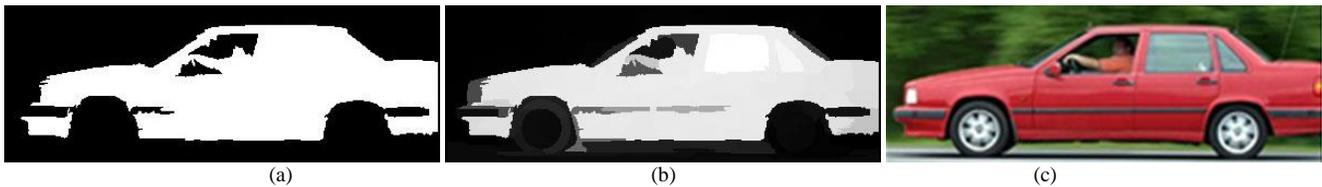


Fig. 8. Salient region extraction, (a): ROI of binary saliency map, (b): ROI of saliency map (c): ROI of original image



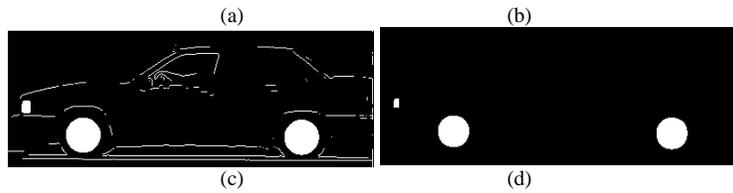


Fig. 9. Particle region segmentation. (a): extracted salient region from original image, (b): thresholding and edge segmentation, (c): region enhancement and filling blobs, and (d): Interested region shape extraction



Fig. 10. A set of particles



Fig. 11. A saliency mask image

TABLE I. COMPONENTS SALIENCY RECOGNITION

**Saliency Components Detection Algorithm**

Input: A set of particles and salient mask.

Extract the edges for all regions and identify the particle properties such as centroid called as  $P_c$  and diameter as  $P_d$ .

Particle mask generation with these parameters: Diameter:  $2P_d$ , Origin: particle center ( $P_c$ )

Generate confluence lines finder (CL) from  $P_c$  to confluence the mask circumference (length of lines are equal to  $2P_d$ )

Measure the distance between points on particle to concern point on salient region touch, using Equation (2).

Count number of equal lines using Equation (3)

Measure the belonging ratio using Equation (4) and identify the particle is belonged to target or not.

Output: a saliency map with its related parts



Fig. 12. Edge segmentation of salient region



Fig. 13. Integration of edge particles and salient region

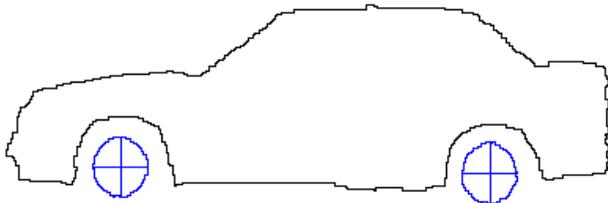


Fig. 14. Diameter of particles

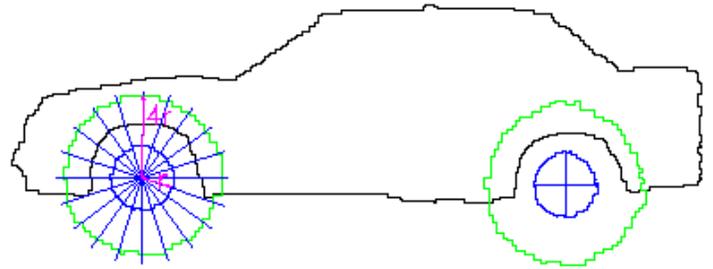


Fig. 18. Sectors and confluence lines (CL) in the mask

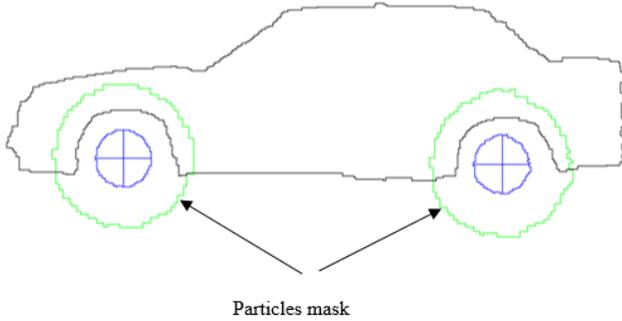


Fig. 15. Generated mask of particles

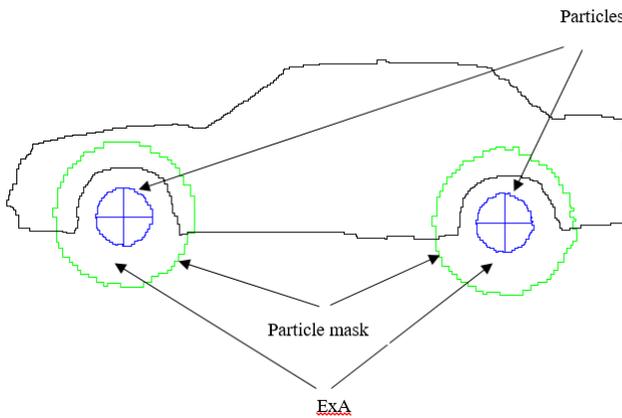


Fig. 16. Exploring area integrated of salient and particles image

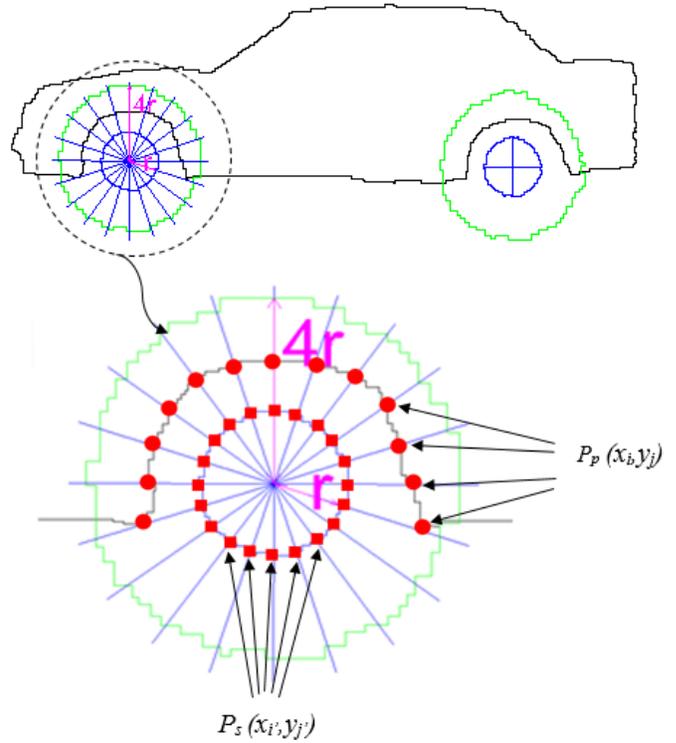


Fig. 19. Points on salient regions and particles

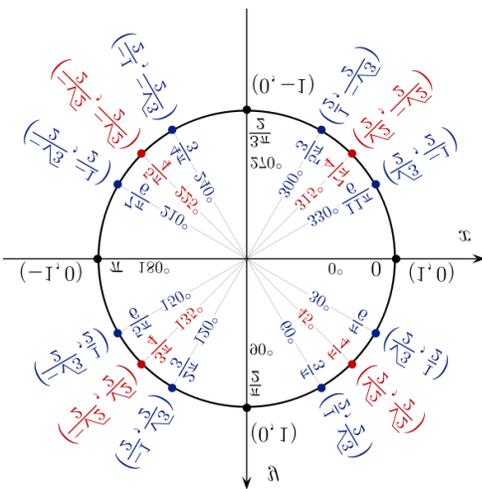


Fig. 17. Circle division into some sectors with angle between CL lines

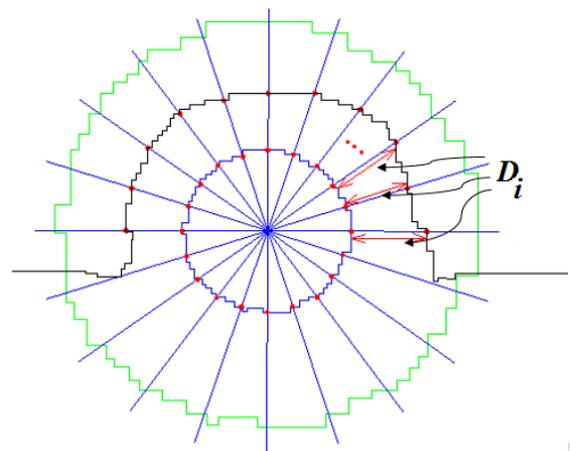


Fig. 20. Distance between salient region ponies and particle points

### III. EXPERIMENTAL RESULTS

In this section, experimental results are presented for the proposed saliency detection method. The proposed method consists of five steps, salient region detection, salient region localization, region of interest extraction, particle segmentation, and multi-part saliency recognition. The salient region detects using the GMR method for an input image. The detected salient region localizes, and the region extract corresponding salient region, is extracted using region of interest extraction. The localization and region of extraction are integrated into one process, because, they are roughly related to each other. The extracted target segments using the segmentation process as explained earlier consist of, an extracted salient region from the original image, there holding and edge segmentation, filling blobs, and shape recognition.

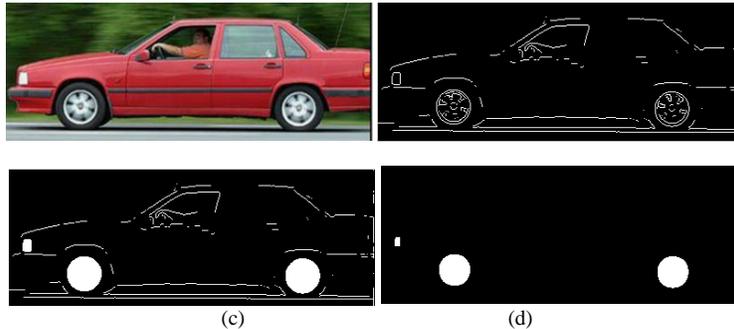


Fig. 21 shows the segmentation process for the salient particle segmentation step.

In the last step of the method, associated particles to the salient region detect using the proposed multi-part saline detection. For these steps, segmented particles (as shown in Fig. 21(a)) from particle segmentation steps are evaluated to check whether they belong to the detected salient region or not. This belonging assessment is based on checking distance measurement between the out-bounding of the segmented particle region and binary salient mask. Finally, those segmented particles belong to salient region are added to the saliency map as shown in Fig. 22(c).

According to the proposed saliency detection method, some image results are illustrated as shown in Fig. 23.

Fig. 21. Particle region segmentation. (a): extracted salient region from original image, (b): thresholding and edge segmentation, (c): filling blobs, and (d): regular shape recognition

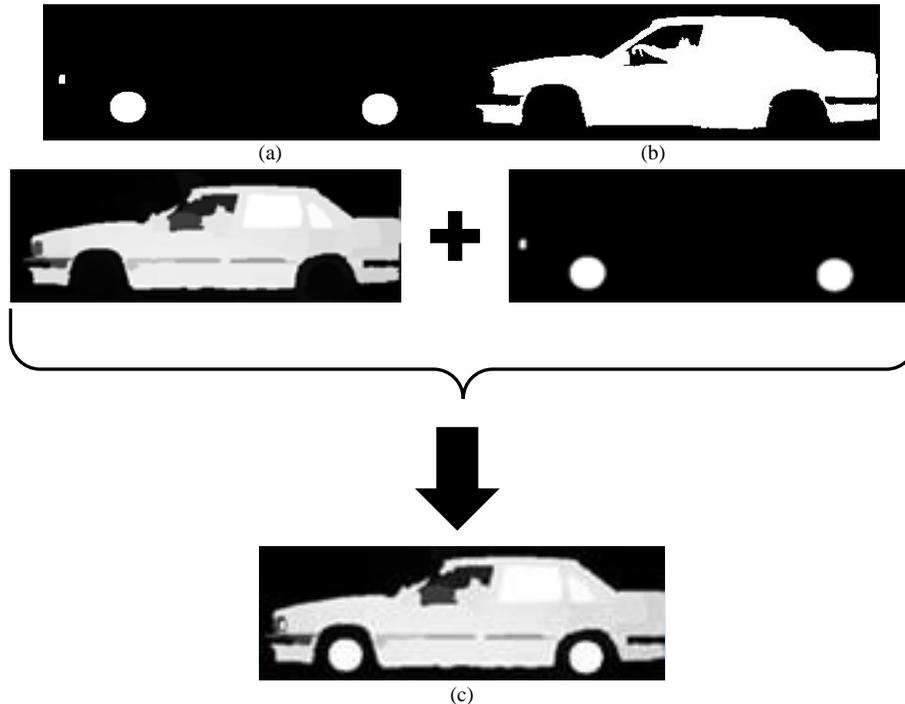


Fig. 22. Multi-part saliency recognition step. (a) segmented particles, (b) binary salient mask

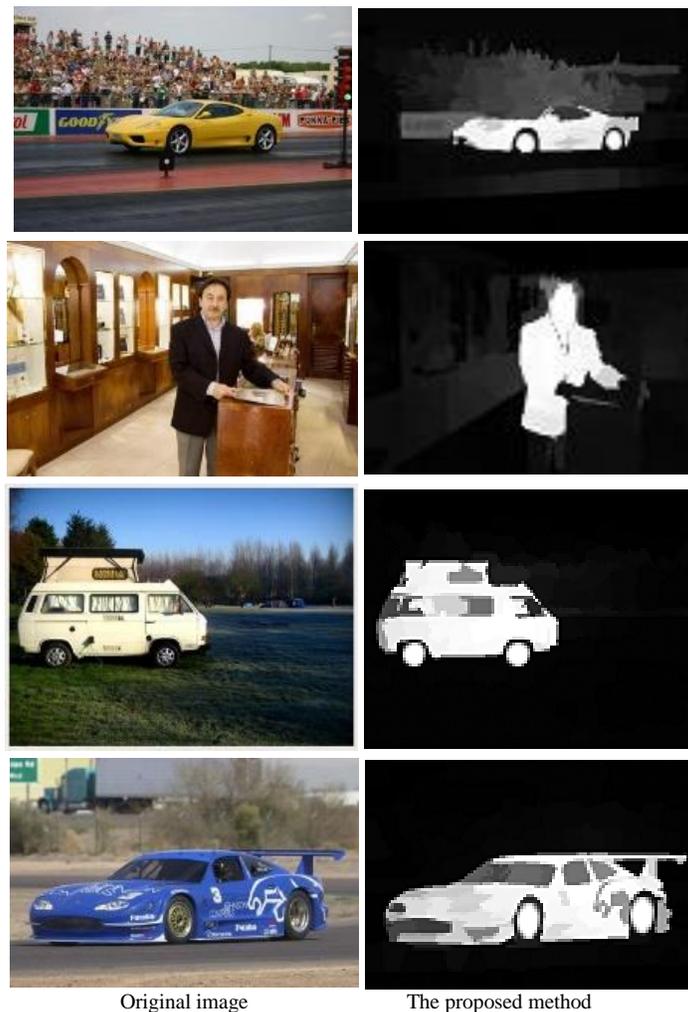


Fig. 23. Experimental result for the proposed saliency detection method

#### IV. CONCLUSION

In this study a new saliency detection method is proposed based on component saliency recognition for a salient region. This component of saliency recognition aims to represent the salient region more accurately because sometimes the salient regions involve several parts. To deal with component saliency recognition for a salient region, the proposed method is structured by five modules, salient region detection, salient region localization, region of interest extraction, particle region segmentation, and possession ratio measurement. In brief, the proposed method is first applied to a graph-based manifold ranking (GMR) method for salient region detection in an image. The detected salient region is localized using centroid finding and region size identification. The localized region is then extracted and called ROI. In ROI, some processes are performed to segment different regions and generate the particles. The BRM is the main module in our saliency detection method. Finally, as experimental results show, the proposed components' saliency recognition for a salient region can represent accurately and efficiently the target appearance representation. For directions of future study, modern advanced methods such as deep learning frameworks can be explored for further improvement of object representation in visual tracking

applications. Moreover, the proposed method can be extended for real time applications.

#### REFERENCES

- [1] Wang, W., et al., Visual saliency detection based on region descriptors and prior knowledge. *Signal Processing: Image Communication*, 2014. 29(3): pp. 424-433.
- [2] Itti, L., C. Koch, and E. Niebur, A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 1998(11): pp. 1254-1259.
- [3] Hou, X. and L. Zhang. Saliency detection: A spectral residual approach. in *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*. 2007. IEEE.
- [4] Garcia-Diaz, A., et al., Saliency from hierarchical adaptation through decorrelation and variance normalization. *Image and Vision Computing*, 2012. 30(1): pp. 51-64.
- [5] Garcia-Diaz, A., et al., On the relationship between optical variability, visual saliency, and eye fixations: A computational approach. *Journal of vision*, 2012. 12(6): pp. 17.
- [6] Achanta, R., et al., Salient region detection and segmentation, in *Computer Vision Systems*. 2008, Springer. pp. 66-75.
- [7] Ming-Ming, C., et al. Global contrast based salient region detection. in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. 2011.

- [8] Goferman, S., L. Zelnik-Manor, and A. Tal, Context-aware saliency detection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 2012. 34(10): pp. 1915-1926.
- [9] Yang, C., et al. Saliency detection via graph-based manifold ranking. in *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*. 2013. IEEE.
- [10] Schauerte, B. and R. Stiefelhagen. How the distribution of salient objects in images influences salient object detection. in *Image Processing (ICIP), 2013 20th IEEE International Conference on*. 2013. IEEE.
- [11] Perazzi, F., et al. Saliency filters: Contrast based filtering for salient region detection. in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. 2012. IEEE.
- [12] Xie, Y., H. Lu, and M.-H. Yang. Bayesian saliency via low and mid level cues. *Image Processing, IEEE Transactions on*, 2013. 22(5): pp. 1689-1698.
- [13] Ali Borji, M.M.C., Huaizu Jiang and Jia Li, Salient Object Detection: A Benchmark *IEEE TRANSACTIONS ON IMAGE PROCESSING*, 2015.
- [14] Chu, M.T. and R.E. Funderlic, The Centroid Decomposition: Relationships between Discrete Variational Decompositions and SVDs. *SIAM J. Matrix Anal. Appl.*, 2001. 23(4): pp. 1025-1044.
- [15] Sideway. Centroid by Geometric Decomposition. 2012; Available from: <http://output.to/sideway/default.asp?qno=120600007>.
- [16] Boston, U.o. Centroid and Area Moments. 2013; Available from: <http://www.bu.edu/moss/mechanics-of-materials-bending-normal-stress/>.
- [17] Ullah I, Jian M, Hussain S, Guo J, Yu H, Wang X, Yin Y. A brief survey of visual saliency detection. *Multimedia Tools and Applications*. 2020. 34605-45.
- [18] Liu Y, Xu Z, Ye W, Zhang Z, Weng S, Chang CC, Tang H. Image neural style transfer with preserving the salient regions. *IEEE Access*. 2019. 40027-37.
- [19] Deeba F, Bui FM, Wahid KA. Computer-aided polyp detection based on image enhancement and saliency-based selection. *Biomedical signal processing and control*. 2020. 101530.
- [20] Jiang B, Lu Y, Lu G, Zhang D. Real noise image adjustment networks for saliency-aware stylistic color retouch. *Knowledge-Based Systems*. 2022.
- [21] Chen B, Cheng Y, Zhang W, Mei G. Investigation on enhanced mathematical morphological operators for bearing fault feature extraction. *ISA transactions*. 2022. 440-59.
- [22] Hu M, Yang J, Ling N, Liu Y, Fan J. Lightweight single image deraining algorithm incorporating visual saliency. *IET Image Processing*. 2022. 16(12). 3190-200.
- [23] Zhu L, Ling H, Wu J, Deng H, Liu J. Saliency pattern detection by ranking structured trees. In *Proceedings of the IEEE International Conference on Computer Vision*. 2017. 5467-5476.
- [24] Bavirisetti DP, Dhuli R. Multi-focus image fusion using multi-scale image decomposition and saliency detection. *Ain Shams Engineering Journal*. 2018. 1103-17.
- [25] Petsiuk V, Jain R, Manjunatha V, Morariu VI, Mehra A, Ordonez V, Saenko K. Black-box explanation of object detectors via saliency maps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021. 11443-11452.

# Building a Machine Learning Powered Chatbot for KSU Blackboard Users

Qubayl Alqahtani, Omer Alrwais

Information Systems, King Saud University, Riyadh, Saudi Arabia

**Abstract**—Chatbots have attracted the interest of many entities within the public and private sectors locally within Saudi Arabia and also globally. Chatbots have many implementations in the education field and can range from enhancing the e-learning experience to answer students' inquiries about course schedules and grades, tracking prerequisites information and elective courses. This work aim is to develop a chatbot engine that helps with frequently asked questions about the Blackboard system, which could be embedded into the Blackboard website. It contains a machine-learning model trained on Arabic datasets. The engine accepts both Arabic textual content as well as English textual content if needed; for commonly used English terminologies. Rasa framework was chosen as the main tool for developing the Blackboard chatbot. The dataset to serve the current need (i.e. Blackboard system) was requested from Blackboard support staff to build the initial dataset and get a sense of the frequently asked questions by KSU Blackboard student users. The dataset is designed to account for as many as possible of KSU Blackboard related inquiries to provide the appropriate answers and reduce the workload of Blackboard system support staff. Testing and evaluating the model was a continuous process before and after the model deployment. The model post-tuning metrics were 93.4%, 92.5%, 92.49% for test accuracy, f1-score and precision, respectively. The average reported accuracy in similar studies were near 90% on average as opposed to results reported here.

**Keywords**—Chatbot; RASA; conversational agents; machine learning

## I. INTRODUCTION

A chatbot is an artificial intelligence (AI) software that can simulate a conversation (or a chat) with a user in natural language through messaging applications, websites, mobile apps or through the telephone [5]. It's an environment that receives questions from users in natural language, relates these questions with a knowledge base, and then answer based on pre-defined answers. Chatbots are more formally referred to in the literature as conversational agents or conversational assistants. The core principle of every conversational agent is to interact with humans using text messages and act as it were able to understand the user and replay with the appropriate message. The origin of computers talking to humans goes back to the start of the computer science field itself. Alan Turing defined a simple test referred to now as the Turing test back in 1950 where a human judge would have to predict if the entity they are communicating with via text is a computer program or not [6]. However, this test's scope is way greater than the case of chatbots, the main difference being that the domain knowledge of a chatbot is narrow compared to the Turing test.

Turing test assumes one can talk about any topic in mind with the agent.

Conversational agent environment consists of five different main parts [7]. Starting with user messages: they are a dynamic input received by the agent to process and replay. They contain a string representation of the actual text sent by the user, and a metadata that contains additional information like a reference to the session the conversation belongs to, and possibly the date and time the message was sent to the agent, on which platform the message was sent from if the agent is linked to more than one message, etc. The agent receives the message along with the information it contains in a read- only mode only with no possible means of making changes to it. The backend is one of the significant parts of the environment that the agent has access to. It contains additional information about the agent users and the database's states to store the user messages, their metadata, and keep track of the conversation events. The agent can view and update certain aspects of the backend. The chatbot can also obtain new information from the user if necessary, by asking the user to provide it.

RASA is a modular design framework proposed by [1] that consists of two main components, Rasa Core for dialog management and Rasa NLU for natural language understanding. Those are open- sourced python libraries for building machine-learning based conversational agents. They provide dialog management and NLU capabilities in an easy manner. By nature, a modular by design architecture allows for easier integration of modules with other systems and services. For instance, Rasa NLU can be used as a service in a different system other than rasa by exposing HTTP APIs for external requests and vice versa for Rasa Core. The code can be found by visiting: <https://github.com/RasaHQ>.

Although chatbots have been present for a long time, 2016, before the spring, was the true start of this technology. There are two main reasons for the renewed interest in chatbots (1) massive advances in artificial intelligence (AI) and (2) a major usage shift from online social networks to mobile applications such as WhatsApp, Telegram, Slack, and many more advances in AI holds a promise that intelligent chatbots are in fact, can be within reach. The increased usage of mobile applications attracts service providers to reach users through them. However, in spite of these advances, chatbot applications entail many challenges that need to be overcome in order to reach desired goals. Chatbots not only imply changes in the interface between users and technology; they imply changing user dynamics and usage patterns. A recent study indicated that 56% of chatbot users were interested in ordering meals from

restaurants using chatbots, while 34% had already ordered at least one meal [2]. Chatbots are considered to be beneficial for retailers in terms of customer service (about 95%), sales/marketing (about 55%), and order processing (about 48%) [3]. Generation Z and Millennials are more interested in using chatbots: 25% of a global sample aged 18 to 34 opted for a personal shopping chatbot [4], and the students using Blackboard system fall under this age range.

King Saud University (KSU) is the largest public university in Saudi Arabia and at any time encompass thousands of students whom some of them use and struggle with the online course and learning management system (blackboard), as such the customer support staff (which is not more than three employees) at KSU is overwhelmed with enquiries which can cause great user dissatisfaction and affect the adaptation of Blackboard system by KSU students. The goal of this work is to build a Minimum Viable Product (MVP) for an Arabic Chatbot which is intended to serve users of Blackboard system from King Saud University students by answering their frequently asked questions about the Blackboard system to reduce the load of answering repeated, one answered questions and allow customer service staff to focus on more dynamic issues that require human intervention.

## II. RELATED WORK

A revision for several chatbot related papers that highlights the usage of chatbots in the education field was conducted, along with other chatbot usages in different areas such as retail and government entities. A review and summarization of those implementations are discussed in the next paragraphs.

### A. Chatbots in Education

A chatbot named EASElective that was built to advise students on what to choose as an elective course was proposed in [22]. EASElective is a conversational agent that was built to supplement existing academic advising systems. It has an interactive, online interface that supports basic official course information to informal students' opinions about that course. Its major components included intent detection, conversational management routines, dialogue design, course information management, and a collection of analyzed students' peers' opinions. In this study, a survey was conducted to capture students' perceptions of the chatbot. The subjects were briefed about the chatbot's purpose and instructed on how to use it and were given up to a half-hour to interact with it and then fill the surveys. The survey results showed that many students preferred to either ask their friends for course information. Around 22% preferred to ask the program leader or use the official university website instead of the chatbot. There were a number of limitations, including the chatbot not having enough interactions to learn from before going live. And also, the chatbot patterns usages are neither recorded nor pre-defined in advance to prepare the appropriate responses.

Another chatbot implementation to enhance the LMS experience was proposed by [23]. This model classifies the main keywords that could be asked by students using R programming language, and this classification is then used in an Artificial Intelligent Markup language (AIML) script as a query. If this query was unsuccessful, it would run against SQL

lite. If neither AIML nor SQL lite worked, then the student query will be transferred to a human agent to take over and answer the query. Although the implementation of AIML scripting language is easy and also free to use as a scripting language, this model is a rule-based model and is less tolerant to changes in users' input and, thus, harder to capture the user intent.

Another study for developing a chatbot for university inquiries was put forward by [24]. This study discussed the development of a deep-learning based chatbot using RASA framework. RASA has many connectors to be used in integrating it with communication platforms. One of them is for Facebook. This chatbot is integrated with FB as the majority population is using FB as their main social media channel. This chatbot uses Long Short-Term Memory (LSTM), which is a recurrent neural network architecture that is used in deep learning. This architecture is included in RASA framework. Although the chatbot performed well in terms of intent classification and provided the appropriate replays, there was a platform limitation as they had to perform platform-specific steps to run the chatbot on Facebook, which can result in some limitations to the interaction with the chatbot.

A chatbot for instantly answering students' questions to reduce teacher's workload was proposed in [25]. It supports multiple common social platforms, including Telegram, Facebook Messenger, and Line. The chatbot can reply to commands and natural language questions. Once the instructors transfer the course-related data to an internet database, the chatbot can reply to questions about the course materials and logistics (e.g., course plan). It also supports student login to provide profile-based answers such as the schedule of student registered courses.

### B. Chatbots in other Fields

Chatbots also have many usages in other fields besides education. Some of those applications are in healthcare, such as self-diagnoses based on symptoms, using chatbots as a communication means in e-commerce websites, providing account data and paying bills in banking, etc. Below are some of the related works of chatbots.

A text-to-text chatbot engages patient's medical issues were proposed by [26]. It's a medical chatbot that diagnoses diseases using AI. This chatbot was built to reduce medical costs and improve patient's accessibility to medical knowledge. In this chatbot, a series of questions about the patient's symptoms are asked to give suggestions that help in clarifying the disease. The accurate disease is found based on the user reply to those series of questions, and in case of major diseases, a doctor is suggested to be consulted. The patient's past responses are recorded, and in order to reach an accurate diagnosis, the patient is asked more specific questions. There are three main components of the system, which are (1) user validation and symptoms extraction from the conversation, (2) mapping of extracted, potentially ambiguous symptoms to their corresponding database codes, and (3) personalized diagnosis and referring the patient to a specialized doctor if required. The sole focus of this system is extracting symptoms by analyzing natural language using NLG components, which in term makes it easier and less technical for the end-user.

Another example of chatbot usage in e-commerce to support customers in their website journey is called “SuperAgent” [27]. This chatbot scrapes public e-commerce websites' content of products description, user questions and answers, and product reviews and feeds them to its knowledge base. It uses NLP techniques to understand users' text and machine learning techniques to predict responses to it, including opinion mining for product reviews, fact QA for product information, and FAQ search for customer reviews and chit-chat for greetings and goodbyes.

ChatPy is one of the chatbot implementations in the wholesale business [28]. “Mundirepuestos” is a wholesaler

automotive spare company. This company is an SME company that started operating in 1992 that specialized in the distribution and sales of Volkswagen, Skoda, and Audi automotive parts. ChatPy is a conversational agent built mainly using a tool called Dialogflow. This tool makes use of intents, actions with parameters, entities, voice-to-text, and text-to-speech with automatic learning. A major reason for choosing this tool was its compatibility with the most known messaging platforms. A summary of chatbot related works in different fields is shown in Table I.

TABLE I. SUMMARY OF RELATED RESEARCH

Title	Field	Method	Limitations
Developing a Chatbot for College Student Programme Advisement [22]	Education	Intent detection, conversational management routines, dialogue design, course information management.	Minimal interactions to learn from before going live. Patterns usages are neither recorded nor pre-defined in advance to prepare the appropriate responses.
Enhancing LMS Experience through AIML Base and Retrieval Base Chatbot using R Language [23]	Education	Artificial Intelligent Markup language (AIML) script as a query, SQLite.	Using a rule-based model only.
Developing FB Chatbot Based on Deep Learning Using RASA Framework for University Enquiries [24]	Education	Long Short Term Memory (LSTM) a recurrent neural network architecture included in RASA framework.	Facebook-specific implementation to run the chatbot which limits customization.
Using a Multiplatform Chatbot as an Online Tutor in a University Course [25]	Education	DialogFlow's ML engine and knowledge base.	Low number of participants for evaluating the model.
A Self-Diagnosis Medical Chatbot Using Artificial Intelligence [26]	Healthcare	Symptom recognition, String Searching Algorithm.	No fallback policy for unknown symptoms.
SuperAgent: A Customer Service Chatbot for E-commerce Websites [27]	E-commerce	Opinion mining for product reviews, fact QA for product information, and FAQ search for customer reviews and chit-chat conversations.	No intent detection.
ChatPy: Conversational agent for SMEs: A case study [28]	Business	DialogFlow's ML engine and knowledge base.	Facebook-specific implementation to run the chatbot which limits customization.

To avoid issues in [22], the system needs to be internally deployed and used by diverse students' backgrounds while recording their usage patterns and interactions which is recorded by default in Rasa framework. Rule-based chatbots like the one presented in study [23] cannot learn which is not the case for Rasa as it has interactive learning capabilities which allows it to learn and refrain from making the same mistake in the future. Unlike the cases presented in study [24] and [28], this work is going to use Rasa API to communicate with the chatbot which remove the platform specific limitations and allows for more customizations. There needs to be a fit number of participants to fairly evaluate the chatbot to overcome the limitation in study [25]. As opposed to the case presented in study [26], Rasa provides a fallback policy that can be triggered when the prediction of the action to be taken is below a specified threshold; this fallback could be used to ask the user to rephrase or show some buttons for the user to choose from. Knowing the user intent can help greatly in providing the right answer to the user question and also helps in performing actions based on the user intent; Rasa uses deep learning embeddings to detect user intention which is not the case in study [27].

III. SYSTEM DESIGN

Fig. 1 below gives an overview of Rasa open-source architecture that consists of two main components which are the Rasa NLU and Dialog management (Core). Rasa NLU is

responsible for predicting intents, extracting entities and retrieving responses. It uses the saved model in the filesystem. The Core component is responsible for choosing the appropriate next action with regards to the conversation context and uses Tracker store to store the conversation states, messages and metadata.

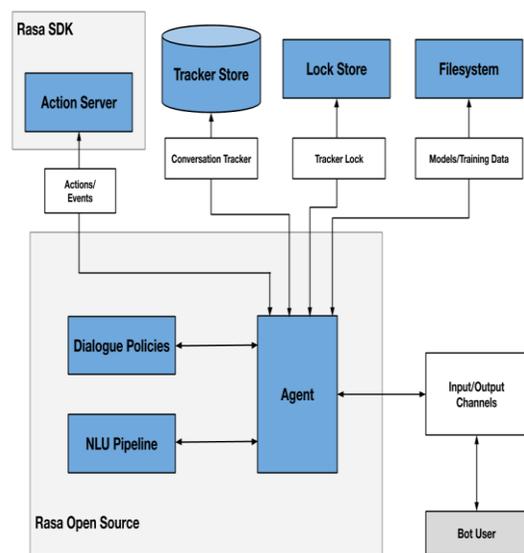


Fig. 1. System architecture overview.

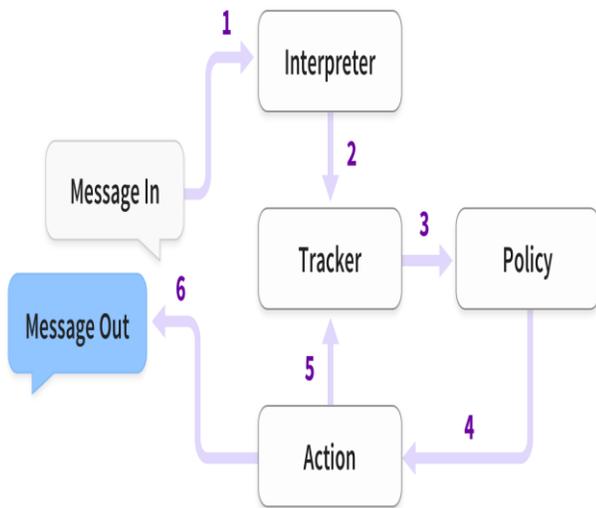


Fig. 2. Rasa architecture.

Rasa ensures that messages are being processed in the right order using the lock store. Actions are running on the Action Server and executed when called by the Core component. Fig. 2 shows the message flow and how Rasa architecture works.

The user first types in a message; this message is then passed to the interpreter in which NLU is used to extract user intention, intent for short, and any entities contained in the text. The conversation state is then saved in a Tracker object, and an event is created, i.e., new message arrival. The state is then received by the policy, and the next action is chosen by the policy to be taken. The action is also logged into the Tracker and then implemented, which could be a response that is based on an external API call or a simple text response that is sent back to the user.

#### A. Data Collection

To gather the required data for training the chatbot to answer most frequent questions asked by students, the LMS Blackboard team admin was contacted to provide this data. This data is in the form of a Word document format and will be used to manually generate the training data and build chatbot stories to train the chatbot. It contains 17 of the frequently asked questions by the Blackboard users. Examples are shown in Appendix 1.

There are two datasets to be built, one for the NLU model which contains examples for each user intent along with labeled entities. The data for this project are in MSA, Modern Standard Arabic and some common English words. The second dataset is for the Core model or the Dialog Management model, it has all the possible flows for the conversation (intents with their corresponding actions). The latter might not be needed when using mapping policy which maps each intent to an action or a template. The dataset to serve the current need (i.e., Blackboard system) was requested from Blackboard support staff to build the initial dataset, and get a sense of the frequently asked questions by KSU Blackboard student users. This data will be then increased by synthesizing text that could

be asked by the chatbot users to increase the chances of understanding students' questions about the Blackboard system. For chatbots systems, the datasets should be continuously updated after deployment for continuous enhancement. The data formats for both NLU and Core are written in a user-friendly format to make it easier to build, revise and edit. For the NLU model, examples for each intent along with labeled entities are created. There are two available formats for building the dataset, either as a json or a markdown format. Markdown formats are the most used as it can be rendered by most text editors.

#### B. Building NLU Corpus

The main goal of building this corpus is to make the chatbot see many examples of what the user might say regarding a specific intention of the user. There are two available formats for building the dataset, either as a json or a markdown format. Markdown formats are the most used one as it can be rendered by most text editors. Below is an example of markdown NLU dataset record.

```
1. ## intent: FAQ_login_issue
2. -(system)[نظام إدارة التعلم] لا أستطيع الدخول على
```

The other format is the json format; it's not sensitive for whitespaces and better in exchanging data among applications. The actual NLU corpus can be found in Appendix 2.

```
1. {
2. "text": " لا أستطيع الدخول على نظام إدارة التعلم",
3. "intent": "FAQ_login_issue",
4. "entities": [{
5. "start": 19,
6. "end": 33,
7. "value": "نظام إدارة التعلم",
8. "entity": "system"
9. }]}
```

#### C. Building Stories

Stories are a type of data that is used in order to teach the chatbot the possible messaging flow with user. Markdowns are used to specify the conversation paths i.e., stories. Below is an example for Dialog management model training data.

```
1. ## story_84854865
2. * greet
3. - utter_ask_howcanhelp
4. - utter_ask_issue
5. * inform("system":"نظام إدارة التعلم")
6. - action_replay_issue
```

The naming convention for stories is to start with two hashes, followed by the story name. Actions are events that start with a dash. The actual Core model corpus can be found in Appendix 3.

#### D. Implementation

Rasa environment requires a list of hard and software requirements for running Rasa on Docker. Although there are minimum hardware requirements on Rasa official website, the hardware requirements depend on the size of the model and training data as the training time and the size of the NLU data are positively correlated. Those requirements need to be met to



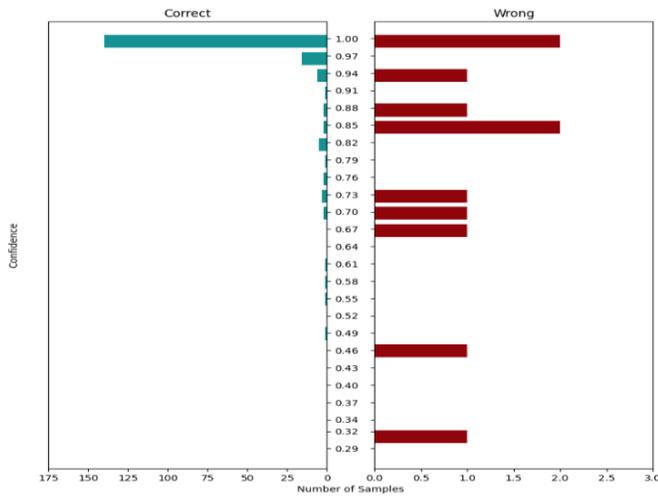


Fig. 5. NLU intent confidence histogram.

From Table III, we can see that all actions were predicted correctly with a value near to 1 for F1-score, precision, and accuracy. The reason for such high results is that the dialog management model is classifying actions based on the results of the intent classifier. If there are no errors in predicting the intention of the user, the prediction of the next action becomes easier and hence, result in a high hit rate.

TABLE III. DIALOG MANAGEMENT RESULTS

Metric	Score
Correct	335/338
F1-score	0.991
Precision	0.992
Accuracy	0.991

As mentioned earlier, we will try to optimize the NLU model by adding more examples and removing examples that might mislead the model into falsely predicting intents. We will also change some of the NLU model configurations to see if those changes yield better results (Table IV).

TABLE IV. PRE-TUNING VS. POST-TUNING METRICS

Metric	Pre-Tuning Score	Post-Tuning Score
Train Accuracy	1	1
Train F1-score	1	1
Train Precision	1	1
Test Accuracy	0.924	0.934
Test F1-score	0.911	0.925
Test Precision	0.922	0.925

Although those are minor changes, they do have an effect and it means that it's possible to further optimize the model by adding more data and tuning the model parameters to find the ones that best fit the data. The average reported accuracy in similar case studies mentioned in [22], [24], [27] is near 90% as opposed to our results which is slightly higher.

## V. CONCLUSION

This work intended to develop a chatbot engine that helps with frequently asked questions about Blackboard system, which could be embedded into Blackboard website. It contains a machine-learning model trained on Arabic datasets. The engine accepts both Arabic textual content as well as English textual content if needed; for commonly used English terminologies. The interactions with the chatbot, as well as the users' evaluations, are stored and used for optimizing the chatbot model to improve future interactions. Chatbot systems development entails many challenges in terms of preparing the training dataset in a way that incorporate as much as possible of users' inquiries without confusion, preprocessing it before feeding it to the NLU model to try to normalize the data and remove unnecessary words and symbols that could confuse the model, and deploy and maintain the model to be used. Rasa framework was chosen to as the main tool for developing the Blackboard chatbot.

The actual chatbot implementation started by preparing the datasets required for Rasa NLU and Core models. The dataset is designed to account for as many as possible of KSU Blackboard- related inquires to provide the appropriate answers and reduce the workload of Blackboard system support staff. When the data was ready, the model training and tuning began along with a number of experimentations to find the best model pipelines that fits the data. The chatbot is built using a combination of tools such as Python for programming, YAML as the markup language.

For future work, the chatbot should be deployed using Docker and Docker-compose for running the chatbot service. The chatbot can also be deployed in a distributed cluster either on cloud or on- premise to handle the workload and make the chatbot system scalable.

## REFERENCES

- [1] T. Bocklisch, J. Faulkner, N. Pawlowski, and A. Nichol, "Rasa: Open source language understanding and dialogue management," arXiv. 2017.
- [2] M. L. ATKINSON, "Voice is the Thing in 2018 - Order automation for restaurants - Orderscape," 2018. <https://orderscape.com/blog/voice-is-thing-2018-order-automation-restaurants/> (accessed Oct. 28, 2020).
- [3] Mindbrowser, "Global Chatbot Trends Report – 2020 - Mindbrowser," 2017. <https://www.mindbrowser.com/chatbot-market-survey-2017/> (accessed Oct. 28, 2020).
- [4] Chatbots Magazine, "Chatbot Report 2019: Global Trends and Analysis | by BRAIN [BRN.AI] CODE FOR EQUITY | Chatbots Magazine," 2019. <https://chatbotmagazine.com/chatbot-report-2019-global-trends-and-analysis-a487afec05b> (accessed Oct. 28, 2020).
- [5] Expert System Team, "Chatbot: What is a Chatbot? Why are Chatbots Important? - Expert.ai | Expert.ai," 2020. <https://www.expert.ai/blog/chatbot/> (accessed Dec. 05, 2020).
- [6] A. M. Turing, "Computing machinery and intelligence," in Machine Intelligence: Perspectives on the Computational Model, 2012.
- [7] A. : Peters, F. Promoteur, and L. Wehenkel, "Master thesis : Design and implementation of a chatbot in the context of customer support." Accessed: Dec. 06, 2020. [Online]. Available: <http://matheo.uliege.be>.
- [8] "(PDF) A Literature Review On Chatbots In Healthcare Domain." [https://www.researchgate.net/publication/334836867\\_A\\_Literature\\_Review\\_On\\_Chatbots\\_In\\_Healthcare\\_Domain](https://www.researchgate.net/publication/334836867_A_Literature_Review_On_Chatbots_In_Healthcare_Domain) (accessed Dec. 06, 2020).
- [9] L. Batista and L. Alexandre, "Text Pre-processing for Lossless Compression," 2008, doi: 10.1109/dcc.2008.78.

[10] Singh and V. Gupta, "Text stemming: Approaches, applications, and challenges," ACM Comput. Surv., 2016, doi: 10.1145/2975608.

[11] A. Oussous, F. Z. Benjelloun, A. A. Lahcen, and S. Belfkih, "ASA: A framework for Arabic sentiment analysis," J. Inf. Sci., 2020, doi: 10.1177/0165551519849516.

[12] N. Lee, K. Kim, and T. Yoon, "Implementation of robot journalism by programming custombot using tokenization and custom tagging," 2017, doi: 10.23919/ICACT.2017.7890154.

[13] T. Jiang, H. Yu, and Y. Jam, "Tibetan word segmentation system based on conditional random fields," in ICSESS 2011 - Proceedings: 2011 IEEE 2nd International Conference on Software Engineering and Service Science, 2011, pp. 446-448, doi: 10.1109/ICSESS.2011.5982349.

[14] J. R. Bellegarda, "Part-of-speech tagging by latent analogy," IEEE J. Sel. Top. Signal Process., 2010, doi: 10.1109/JSTSP.2010.2075970.

[15] L. Yang, M. Zhang, Y. Liu, M. Sun, N. Yu, and G. Fu, "Joint POS Tagging and Dependence Parsing with Transition-Based Neural Networks," IEEE/ACM Trans. Audio Speech Lang. Process., 2018, doi: 10.1109/TASLP.2017.2788181.

[16] B. Chen and D. Ji, "Chinese semantic parsing based on dependency graph and feature structure," 2011, doi: 10.1109/EMEIT.2011.6023005.

[17] Z. Li, M. Zhang, W. Che, T. Liu, and W. Chen, "Joint optimization for Chinese POS tagging and dependency parsing," IEEE Trans. Audio, Speech Lang. Process., 2014, doi: 10.1109/TASLP.2013.2288081.

[18] L. Gao and H. Chen, "An automatic extraction method based on synonym dictionary for web reptile question and answer," 2018, doi: 10.1109/ICIEA.2018.8397745.

[19] S. Qin, J. Song, P. Zhang, and Y. Tan, "Feature selection for text classification based on part of speech filter and synonym merge," 2016, doi: 10.1109/FSKD.2015.7382024.

[20] N. Chetty, K. S. Vaisla, and N. Patil, "An Improved Method for Disease Prediction Using Fuzzy Approach," 2015, doi: 10.1109/ICACCE.2015.67.

[21] B. Xu, Y. Ma, and H. Lin, "A hybrid deep neural network model for query intent classification," J. Intell. Fuzzy Syst., 2019, doi: 10.3233/JIFS-182682.

[22] C. Chun Ho, H. L. Lee, W. K. Lo, and K. F. A. Lui, "Developing a Chatbot for College Student Programme Advise," 2018, doi: 10.1109/ISET.2018.00021.

[23] V. K. Shukla and A. Verma, "Enhancing LMS Experience through AIML Base and Retrieval Base Chatbot using R Language," 2019, doi: 10.1109/ICACTM.2019.8776684.

[24] Y. Windiatmoko, A. F. Hidayatullah, and R. Rahmadi, "Developing FB Chatbot Based on Deep Learning Using RASA Framework for University Enquiries," Sep. 2020, Accessed: Nov. 22, 2020. [Online]. Available: <http://arxiv.org/abs/2009.12341>.

[25] L. K. Lee, Y. C. Fung, Y. W. Pun, K. K. Wong, M. T. Y. Yu, and N. I. Wu, "Using a Multiplatform Chatbot as an Online Tutor in a University Course," 2020, doi: 10.1109/ISET49818.2020.00021.

[26] S. Divya, V. Indumathi, S. Ishwarya, M. Priyasankari, and S. Kalpana Devi, "A Self-Diagnosis Medical Chatbot Using Artificial Intelligence," J. Web Dev. Web Des., 2018.

[27] L. Cui, S. Huang, F. Wei, C. Tan, C. Duan, and M. Zhou, "Superagent: A customer service chatbot for E-commerce websites," 2017, doi: 10.18653/v1/P17-4017.

[28] P. Perez, F. De-La-Cruz, X. Gueron, G. Conrado, P. Quiroz-Palma, and W. Molina, "ChatPy: Conversational agent for SMEs: A case study," in Iberian Conference on Information Systems and Technologies, CISTI, Jun. 2019, vol. 2019-June, doi: 10.23919/CISTI.2019.8760624.

[29] "Docker Compose Installation." <https://rasa.com/docs/rasa-x/installation-and-setup/install/docker-compose/#docker-compose-manual-install> (accessed Apr. 02, 2021).

[30] "What is Python? Executive Summary | Python.org." <https://www.python.org/doc/essays/blurb/> (accessed Apr. 02, 2021).

APPENDICES

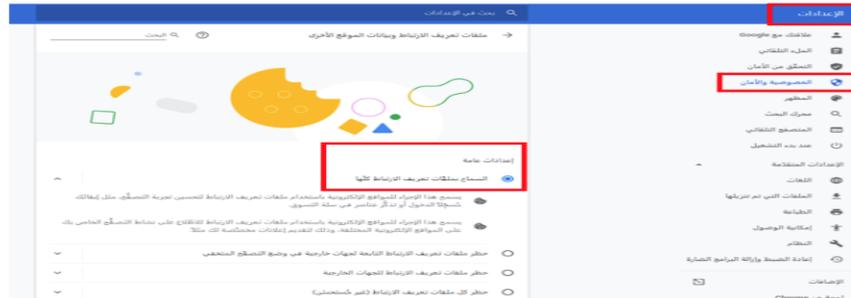
Appendix 1: Raw Data: Below is a snippet of the raw data used in building the chatbot. For the full file please download it through this link: <https://drive.google.com/drive/folders/1zKxniGzjqYUCYAcHgUM1I8o-Ma-sg8Ag?usp=sharing>.

**أسئلة الطلاب الشائعة لنظام إدارة التعلم LMS**

س/ يوجد لدي مشكلة بالصوت داخل جلسة الفصل الافتراضي؟  
هم بتفعيل خيار الوصول إلى الميكروفون من المتصفح المستخدم ، وذلك من خلال النقر على علامة الففل بجانب اسم الموقع بالأعلى ثم إعدادات الموقع ثم اختر السماح للميكروفون

س/ عند الدخول على نظام إدارة التعلم يوجد لدي مشكلة بالصوت؟  
لا تستخدم المتصفح من خلال الحوالم، بل عليك أن تقوم بتحميل تطبيقين على حوالم أو الجهاز اللوحي وهما:  
Blackboard -1  
Collaborate -2

س/ عند الضغط على قائمة الفصول الافتراضية يظهر شاشة سوداء ولا تظهر لي الجلسات؟  
اذهب إلى علامة الففل lock بجانب اسم الموقع بالأعلى ثم إعدادات الموقع ثم الخصوصية والأمان ثم فعل السماح بملفات تعريف الارتباط كلها.



س/ كيف أصل إلى تسجيلات المحاضرات السابقة؟  
اذهب إلى الفصول الافتراضية ثم علامة الإعدادات ثم انقر على التسجيلات واتبع الخطوات.



Appendix 2: NLU Corpus: Below is a snippet of the NLU data used in building the chatbot intent classifier. For the full file please download it through this link: <https://drive.google.com/drive/folders/1TN3yluifFYDXm7FNwdNn-AbthC70hWYg?usp=sharing>.

```
- intent: greet
examples: |
- السلام عليكم
- السلام عليكم ورحمة الله وبركاته
- هلا اخوي
- الله يحبك
- اهلا وسهلا
- حياك الله ياأخي
- سلم عليكم
- سلام
- هلا
- أهلا
- صباح الخير
- مساء الخير
- صباحك الله بالخير
- مساءك الله بالخير
- هلا اخوي
- ايدا
- هلا والله
- صباحالخير
- مساءالخير
- يسعد مسائلك
- يسعد صباحك

- intent: FAQ_how_to_see_screen_recordings
examples: |
- كيف اشوف تسجيل المحاضرات
- كيف اشوف المحاضرات السابقة
- وين القى المحاضرات السابقة
- وين القى المحاضرات الماضية
- كيف اشوف الجلسات المسجلة
- الدكتور كان يسجل المحاضره كيف
- وين القى الجلسات المسجلة
- وين احصل تسجيل المحاضرات
- ماعرفت اطلع تسجيل المحاضرات
```

Appendix 3: Core Model Corpus: Below is a snippet of the Core (Dialog) data used in building the chatbot Dialog model. For the full file please download it through this link: <https://drive.google.com/drive/folders/1uzF31NWfO6sAv5VzIAkcrUD6hIv0c-Y8?usp=sharing>.

```
- story: FAQ_in_lms_sound_issue path
steps:
- intent: greet
- action: utter_greet
- action: action_listen
- intent: FAQ_in_lms_sound_issue
- action: utter_FAQ_in_lms_sound_issue
- action: action_listen
- action: utter_goodbye

- story: FAQ_how_to_change_password path
steps:
- intent: greet
- action: utter_greet
- action: action_listen
- intent: FAQ_how_to_change_password
- action: utter_FAQ_how_to_change_password
- action: action_listen
- action: utter_goodbye

- story: FAQ_how_to_see_alerts_in_lms path
steps:
- intent: greet
- action: utter_greet
- action: action_listen
- intent: FAQ_how_to_see_alerts_in_lms
- action: utter_FAQ_how_to_see_alerts_in_lms
- action: action_listen
- action: utter_goodbye
```

Appendix 4: Domain File: Below is a snippet of the Domain file used in building the chatbot. For the full file please download it through this link:  
<https://drive.google.com/drive/folders/1GPqHIRBFg9slW5ddSGUZhQqxIIERYzvK?usp=sharing>.

```
intents:
- greet:
  use_entities: true
- goodbye:
  use_entities: true
- at_fpm:
  use_entities: true
- help:
  use_entities: true
- FAQ_in_lms_sound_issue:
  use_entities: true
- FAQ_no_session_black_screen:
  use_entities: true
- FAQ_how_to_see_green_recordings:
  use_entities: true
- FAQ_changes_in_edu_gate_didnt_appear_in_lms:
  use_entities: true
- FAQ_courses_in_lms_are_not_sorted:
  use_entities: true
- FAQ_how_to_change_password:
  use_entities: true
- FAQ_cant_login_to_lms:
  use_entities: true
- FAQ_how_to_see_alerts_in_lms:
  use_entities: true
- FAQ_how_to_resubmit_document:
  use_entities: true
- FAQ_some_options_doesnt_appear:
  use_entities: true
- FAQ_courses_doesnt_appear_in_mobile:
  use_entities: true
- FAQ_saudi_digital_library_login_issue:
  use_entities: true
- FAQ_virtual_sessions_doesnt_appear_in_side_panel:
  use_entities: true
- FAQ_im_the_only_one_in_session:
  use_entities: true
- FAQ_my_submission_is_pending:
  use_entities: true
- FAQ_submit_button_is_not_working:
  use_entities: true

- bot_challenge:
  use_entities: true
entities: []
slots: {}
responses:
utter_greet:
- text: أهلا عزيزي الطالب، كيف أقدر أخدمك؟
utter_did_that_help:
- text: هل تمت خدمتك بشكل صحيح؟
utter_goodbye:
- text: إلى اللقاء!
utter_iamabot:
- text: من مطلاب جامعة الملك سعود LMS أنا بوت مطور لخدمة مستخدمي
utter_FAQ_in_lms_sound_issue:
- image: https://i.ibb.co/QnffkXz/Screen-Shot-2021-04-05-at-9-48-57-AM.png
text:
قم بتحميل خيار الوصول إلى الميكروفون من المتصفح المستخدم ، وذلك من خلال النقر على علامة القفل بجانب اسم الموقع بالأعلى ثم إعدادات الموقع ثم
اختر السماح للميكروفون
: لا تستخدم المتصفح من خلال الجوال، بل عليك أن تقوم بتحميل تطبيقين على جوالك أو الجهاز اللوحي وهما
Blackboard
Collaborate
utter_FAQ_no_session_black_screen:
- image: https://i.ibb.co/1rpgqxy/no-session.png
text: بجانب اسم الموقع بالأعلى ثم إعدادات الموقع ثم الخصوصية والأمان ثم فعل السماح بملفات تعريف الارتباط كلها lock اذهب إلى علامة القفل.
```

Appendix 5: Test Stories: Below is a snippet of the Test Stories used for testing the chatbot. For the full file please download it through this link:  
[https://drive.google.com/drive/folders/1WtU-\\_ph30Kkjf2ycC2tpVRx1843Y0puh?usp=sharing](https://drive.google.com/drive/folders/1WtU-_ph30Kkjf2ycC2tpVRx1843Y0puh?usp=sharing).

```
stories:
- story: FAQ_in_lms_sound_issue test story 1
steps:
- user: |
أهلا
intent: greet
- action: utter_greet
- action: action_listen
- user: |
عندي مشكلة فالصوت ما ادري وش السبب
intent: FAQ_in_lms_sound_issue
- action: utter_FAQ_in_lms_sound_issue
- action: action_listen
- user: |
شكرا
intent: goodbye
- action: utter_goodbye

- story: FAQ_cant_login_to_lms test story 2
steps:
- user: |
سلام عليكم
intent: greet
- action: utter_greet
- user: |
lms ما اقدر ادخل على ال
intent: FAQ_cant_login_to_lms
- action: utter_FAQ_cant_login_to_lms
- action: action_listen
- user: |
يعطيك الف عافيه
intent: goodbye
- action: utter_goodbye
```

# WEB-based Collaborative Platform for College English Teaching

Yuwan Zhang

Henan Polytechnic Institute, School of Culture, Tourism and International Education, Nanyang, 473000, China

**Abstract**—At present, colleges and universities are trying to apply online education. The online college English course teaching cooperation platform is an important part of college English teaching. At present, teachers' scoring method for students' online examination on this kind of platform is mainly human scoring, which has a low efficiency. In view of this, based on the characteristics of web, this paper constructs an English test paper scoring algorithm based on text matching degree algorithm and improved KNN algorithm. The data analysis type of the algorithm is mainly prescriptive analysis that is, judging whether to give points according to the characteristics of the data. The automation and high efficiency of the algorithm can save a lot of human costs in the field of online education. The experimental results show that the recall rate of the improved KNN scoring algorithm for specific semantic topics is up to 0.9, and only 7.3% of students report that the algorithm misjudges their grades. The results indicate that the algorithm has the potential to be applied to the Web-based college English course teaching collaboration platform and reduce the workload of teachers and improve their efficiency.

**Keywords**—Web; text similarity; KNN algorithm; vocabulary matching; network teaching

## I INTRODUCTION

With the development of network and information technology, online teaching mode is gradually adopted by major universities. This mode is different from traditional teaching, and teaching methods are more diversified and convenient for students [1]. English is a required course for most majors in colleges and universities, and the number of users of its web-based teaching collaboration platform has been very large. Therefore, it often takes more time for English teachers to correct students' test papers online, which increases labor costs and reduces efficiency [2]. The reason why the English test paper correction of the web network teaching collaboration platform cannot be fully automated is that it is difficult to judge the compliance of students' answers with the standard answers through algorithms [3]. Among related technologies, text similarity detection is a technology to calculate the same degree of two texts, and k-nearest neighbor (KNN) algorithm is a mature classification algorithm [4]. In order to solve the automation problem of online English test paper marking, this paper studies the scoring algorithm of English test paper on the WEB online teaching platform based on these two technologies. The goal of the algorithm is to provide an automatic marking method, which can complete the marking of objective and subjective questions with high accuracy.

The article is divided into five parts. The second part is related works, which describes the latest progress in research related fields and shows the basis of research. The third part is the method, which describes the construction and related settings of the algorithm. The fourth part is the experiment, which describes the performance test of the algorithm. The fifth part is conclusion, which summarizes the whole research and proposes the future direction of the research.

## II RELATED WORKS

In recent years, relevant literature and research results have been sorted out. The fields involved are mainly the latest development of online education research, as well as the application of KNN and text similarity technology. Some studies have explored the web-based online teaching platform and related technologies. Poultsakis led his team to study the application of digital learning and related tools in Greece and found that the popularity of digital learning is very low [5]. This is largely due to the backwardness of digital learning-related technologies in the region, which leads teachers to believe that the teaching effect of digital learning is poor and do not trust digital learning. Stamatios Papadakis et al. studied the situation of students using mobile phones to access a learning management system [6]. According to the survey results, there are differences in the use of the learning management system by students through mobile phones. Due to the limitations of reliability and practicality, the system is more used by students as a document library than a learning tool. Panagiotakopoulos and his team proposed a structured approach to develop an outreach plan aimed at improving the coding ability of pre-service and in-service teachers [7]. The project is a successful online teaching plan, with the actual number of classroom logins and completion rate of 70.84%. Researchers believe that this is because the design of the project is easy to use. Christianson designed a remote online voting system to help students enhance their sense of participation. The students said they had a positive experience in this way of participation [8]. Karakose and his partners studied the psychological state and Internet addiction of school administrators and teachers under the background of the epidemic, and the results showed that Internet addiction indirectly affected teachers' loneliness and happiness [9]. The research expenditure on teachers' mental health also needs attention, and it is one of the feasible schemes to reduce teachers' workload through innovative algorithms. Lavidas K et al. studied the online teaching of preschool teachers during the epidemic, and pointed out that preschool teachers use less online teaching, and they prefer real communication framework and teaching process [10]. Katsaris and his partner

analyzed 42 papers related to online teaching from 2015 to 2020, introduced the theoretical and technical background of adaptive e-learning system, and emphasized the importance and efficiency of using learning style in adaptive learning process [11]. From the conclusion of this article, we can see that there is little research on automatic algorithms for marking test papers in online English teaching. From the research in the field of online education, we can see that the research of auxiliary technology for teacher's examination paper marking has received little attention and there is still a lot of research space.

Some researchers have also made corresponding explorations in KNN and text similarity detection. Zardari Za et al. have developed a detection and prevention algorithm to deal with network attacks. The algorithm is based on KNN and can distinguish abnormal nodes from normal nodes according to their behavior differences. Experiments show that the algorithm can effectively reduce latency and increase network throughput [12]. Wang and his team proposed a weighted KNN algorithm, which is calculated based on signal similarity and spatial location. They applied it to fingerprint location. The evaluation results show that the algorithm can improve the accuracy of fingerprint location [13]. In order to solve the problem of abnormal bridge health monitoring data, Lei Z and his research team proposed a KNN based bridge health monitoring algorithm. The algorithm measures the pattern distance between time subsequences according to the similarity of time series, and then selects abnormal patterns. The experimental results show that the model has certain reference value [14]. Pang and his collaborators put forward a Chinese text similarity detection method based on the semantics of feature phrases, which obtains feature phrases by replacing concepts and calculates text similarity. Experiments show that the output results are reliable [15]. Yang et al. proposed a news topic text detection method based on capsule semantic graph, which has lower time complexity than traditional detection, and the experimental data show that it has high accuracy and recall [16]. Franclinton and his research team proposed an extensible code similarity detection model with online architecture rather than local spikes. The experimental results show that the model can better maintain the academic integrity in programming [17].

Through combing the research trends in related fields, it is found that the web-based online teaching platform has been widely practiced and applied, but there is a lack of research on automatic scoring of English test papers. On the other hand, KNN and text similarity detection technology have also been applied in many fields. The research combines these two technologies in order to make contributions to the research of automatic scoring technology in the web online teaching platform.

### III CONSTRUCTION OF SCORING ALGORITHM FOR TEST PAPER OF WEB-BASED COLLABORATIVE TEACHING PLATFORM FOR COLLEGE ENGLISH COURSE

#### A. Web Architecture Selection and Vocabulary Matching Algorithm Construction

The WEB-based College English course network teaching cooperation platform uses B/S architecture to send documents. The structural diagram of B/S architecture is shown in Fig. 1. This architecture has very low requirements on the equipment of the client layer. Only a normal WEB browser is required to participate in College Online English courses. Due to the great differences in the electronic equipment used by college students, it is inevitable to make mistakes when the client mode. Therefore, the B/S architecture is the most secure [18]. In addition, the architecture has good reusability and scalability, which is conducive to the long-term use and version update of the English course online education collaboration platform [19].

After the framework of the platform is determined, the corresponding algorithm can be built. The test paper questions adopted by the WEB-based College English course network teaching and writing platform can be divided into objective questions and subjective questions. Due to the existence of standard answers to objective questions, students need to be completely consistent with the standard answers to score. Therefore, the idea of complete vocabulary matching can be used to build an objective question scoring algorithm. The judgment formula is shown in formula (1).

$$S = \frac{S_t N}{N_t} \quad (1)$$

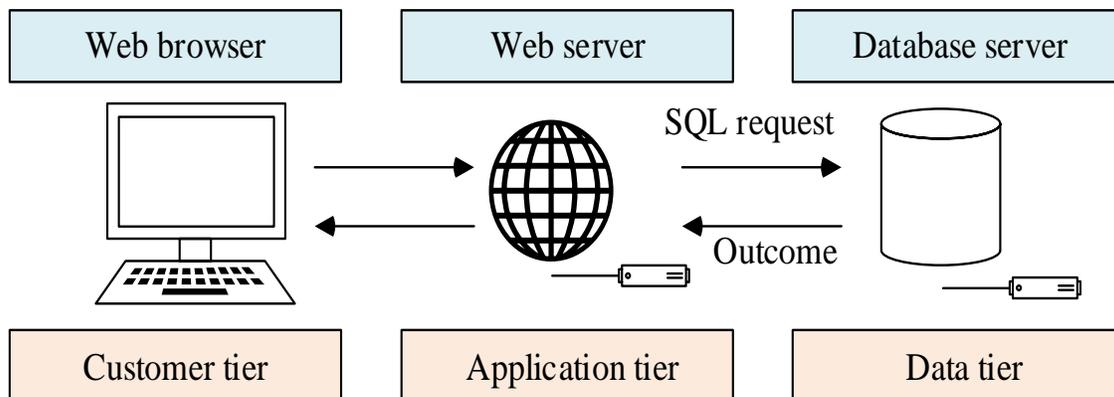


Fig. 1. B/S framework principle.

In formula (1),  $S$  represents the student's score on the question, and  $N$  represents the number of matching keywords with the standard answer, while  $S_t$  and  $N_t$  respectively represent the total score of the question and the number of all keywords. Compared with the scoring of objective questions, the algorithm of subjective questions is more complex. The reason is that both the reference answers of subjective questions and the answers of students are presented in the form of paragraphs. At the same time, the logic adopted by the two texts is not necessarily the same. There will also be differences in the keywords used by the two. It is common for the reference answers and the keywords used by students to be synonyms or superior and subordinate words (Liu M et al. 2021) [20]. The method that teachers use for manual marking of subjective questions is usually judged according to the coincidence degree of key words and reference answers in students' texts, as well as their writing logic and the purity of the overall content (Khan I u et al. 2021) [21]. The scoring algorithm design of English subjective questions on the web online teaching platform refers to the logical design of teachers' scoring, and it uses the method of prescriptive analysis to evaluate the score. Its process is shown in Fig. 2.

According to Fig. 1, the scoring logic of the algorithm is an imitation of the teacher's human scoring. On the one hand, the evaluation matches the keyword of the standard answer with the text of the student's answer, and gives the score according to the proportion of the number of successful matches in the total number. On the other hand, the text similarity between the standard answer and the student's answer is calculated, and the score is given according to the degree of fit between the two texts, and then the two scores are combined according to a certain weight to obtain the final score. In this model, the word matching algorithm adopts a two-way matching algorithm.

This algorithm is an optimization of the ordinary single item matching algorithm. It can distinguish keywords from the forward and reverse directions. In this algorithm, the matching degree between a keyword and the keyword in the student's answer is calculated by the common formula (2).

$$\alpha(K_i, D_i) = \frac{\text{Max}(K_i, K_j)}{m_i} \quad (2)$$

In equation (2),  $\alpha(K_i, D_i)$  is the ratio of a keyword to the number of characters of the current keyword. When the value is greater than the given threshold, the keyword matching is considered successful, otherwise, the matching is considered

failed.  $K_i$  and  $K_j$  are respectively the number of characters of keywords in the student text when the forward

matching and reverse matching are successful, while  $m_i$  is the number of characters of keywords in the reference answer. This algorithm effectively avoids the recognition failure due to the difference between the students' words and the reference answer. After the keyword matching condition is obtained, the keyword score of the question can be calculated. The calculation logic of the score of the subjective question is similar to that of the objective question. It is judged by the ratio of the total number of identified successful keywords to the total number of keywords in the reference answer. The calculation process is shown in formula (3).

$$S = S_t \frac{\sum_{i=1}^{N_t} \alpha(K_i, D_i)}{N_t} \quad (3)$$

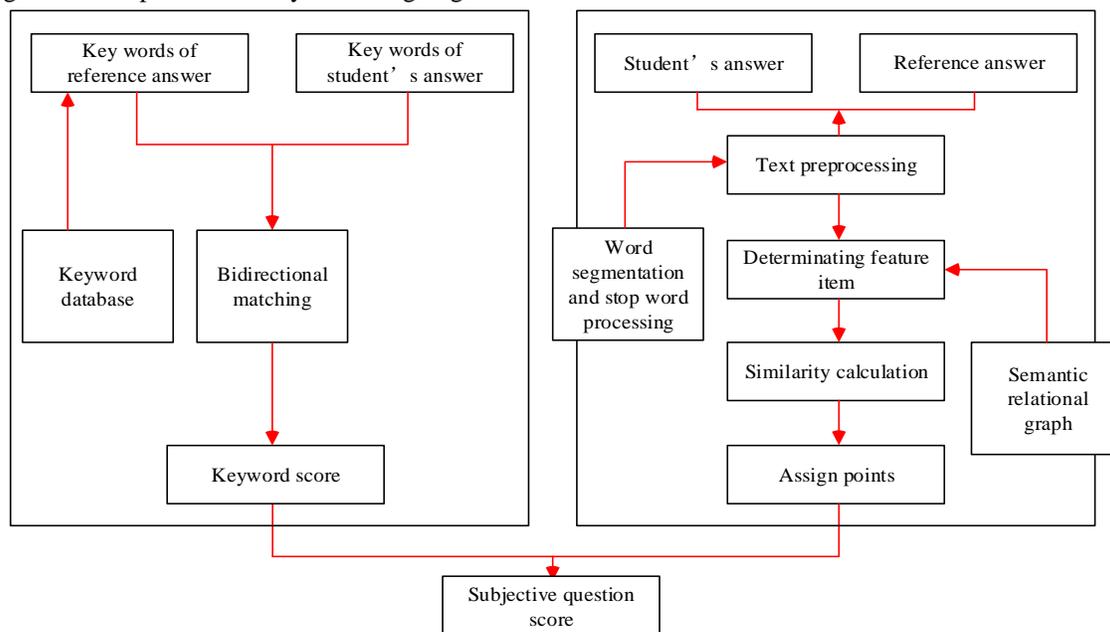


Fig. 2. Flow chart of subjective question scoring algorithm.

**B. Text Similarity Detection Algorithm**

According to the flow chart of subjective question scoring algorithm, the flow of text similarity detection algorithm is to preprocess student answers and reference answers, match feature vectors by combining semantic association, then calculate similarity, and finally calculate scores according to the closeness of answers. Text preprocessing mainly includes two steps: word segmentation and stop word filtering. The word segmentation tool uses THULAC as the word separator. The tool has high word segmentation accuracy and good recognition ability for professional terms. At the same time, it has good adaptability to the web [22]. Stop word filtering is the operation of filtering words such as “very” and “do” that have little effect on the actual meaning of the text, which can effectively reduce the workload of subsequent recognition and matching, save computing resources and improve speed [23]. In the subsequent feature item determination steps, the traditional feature item weight calculation does not consider the semantic problem, but the proposed vector space model takes the semantics into account when selecting feature items, so it is necessary to build a semantic association diagram. In this study, the semantic association diagram is made based on the How Net semantic knowledge dictionary, and its principle is shown in Fig. 3.

In Fig. 2,  $T_1, T_2, T_3$  and other items are semantic topic nodes. These items are a large number of semantic hypernyms, but these items do not have hypernyms. “Sports”, “biology” and other words can be used as semantic topics, while  $t_4, t_5, t_6$  and other items are called non semantic topic nodes. These nodes belong to the hyponymy of one or more semantic topics,

and may have their own hypernymic or hyponymic words, “Basketball” is the non-semantic topic node of “Sports”. Based on the upper and lower semantic relations of the semantic association graph, the upper semantic relations of the semantic association graph can be expressed in mathematical form, and the expression is shown in formula (4).

$$\begin{cases} L_U(t_i) = L(t_i), U = 1 \\ L_U(t_i) = \cup_{t_k \in L_{U-1}(t_i)} L(t_k), U > 1 \end{cases} \quad (4)$$

In formula (4),  $L_U(t_i)$  represents the set of semantics starting from  $t_i$  and going up the  $U$  layer, besides  $t_k$  and  $t_i$  respectively represent different semantic nodes. Based on this formula, the union of all the superscript nodes of any node can be obtained, that is, the set of the node. The expression for finding the set is shown in formula (5).

$$L_U(t_i) = L1(t_i) \cup L2(t_i) \cup L3(t_i) \cup \dots \quad (5)$$

After the definition of semantic association graph is completed, it is necessary to build a semantic space vector model. The model is set as  $R$ , its dimension is set as  $D$ , and the feature vector is  $\vec{t}$ . Then the expression of the model and semantic feature vector is shown in equation (6).

$$\begin{cases} R = T_{\geq 0}^D \\ \vec{t} = (t_1, t_2, \dots, t_n) \end{cases} \quad (6)$$

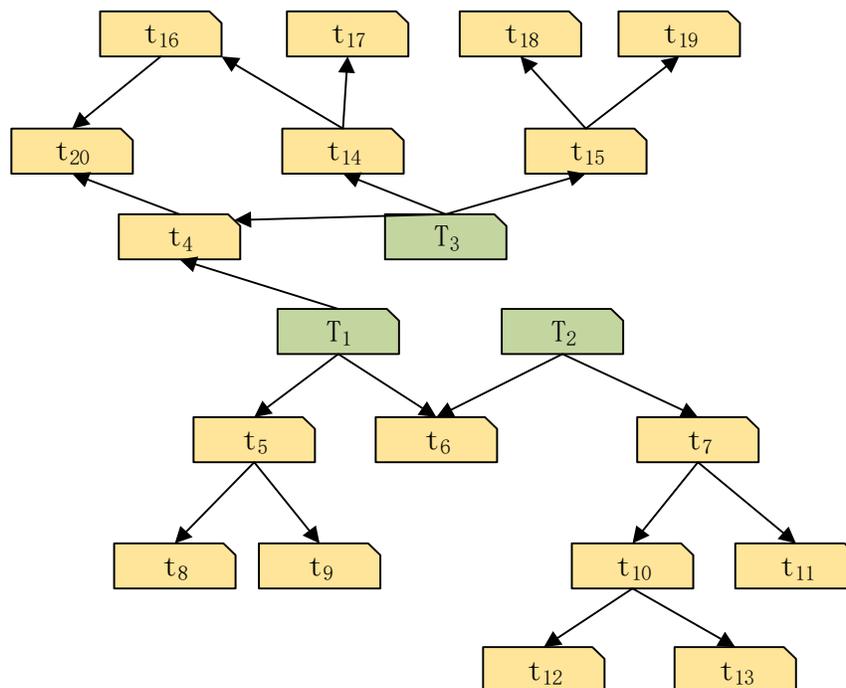


Fig. 3. Schematic diagram of semantic association.

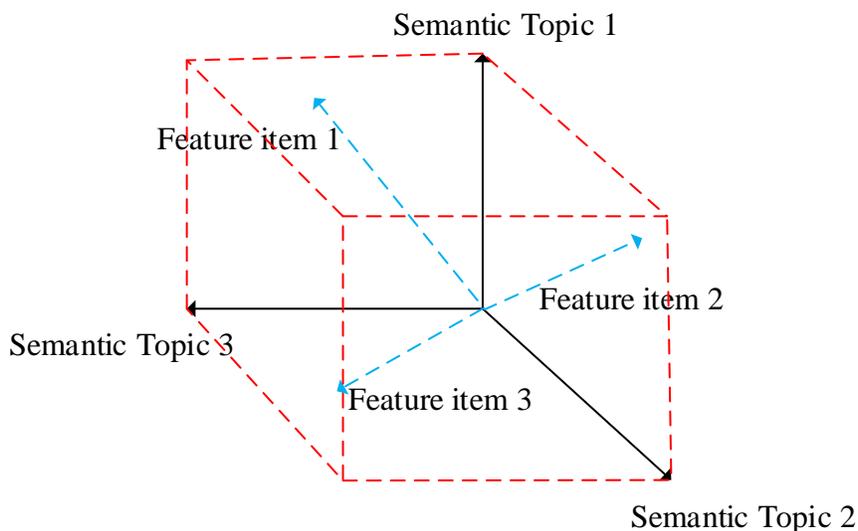


Fig. 4. Schematic diagram of feature item and semantic topic space.

The number of  $t_i$  in equation (6) is determined by the dimension, that is,  $i \in \{1, 2, \dots, D\}$ , and  $t_i \in [0, 1]$ . According to this formula, the schematic diagram of the semantic space vector model is shown in Fig. 4. Each coordinate axis in the figure represents a semantic topic. The more a feature item matches a semantic topic, the greater its value on the coordinate axis of the topic. If a feature item is related to a plurality of semantic topics, its vector will be between the two fields. The correlation between each feature item and each semantic topic depends on the weight of each vector. The higher the correlation with a topic, the higher the weight of its corresponding component.

After defining the semantic space vector model, it is necessary to quantify and express the semantic feature vector in an appropriate way. Quantification is to meet the needs of text similarity calculation, and the appropriate expression can simplify the calculation and improve the efficiency of the algorithm. In terms of quantification, the following rules are designed. First, the weight of the feature item ranges from 0 to 1. The larger the value, the better the feature item reflects the semantics. When the weight is close to 1, the feature item is considered to be basically equivalent to the semantic topic. When the weight is close to 0, the feature item is considered to be basically irrelevant to the semantic topic. Secondly, in terms of the angle of feature items, it is stipulated that the angle between synonyms and feature items not in any semantic field is 0, the angle between synonyms and hyponyms should be close to 0, and the vector angle between feature items in different fields is 90 degrees. In the aspect of feature vector representation, the occurrence times of feature items in the text and their weights in the semantic space vector model are used as variables to represent the feature vector, and the expression is shown in formula (7).

$$|\alpha| = \left| \sum_{i=1}^n F(t_i) \vec{t}_i \right| \quad (7)$$

In formula (7),  $|\alpha|$  is the feature item vector after normalization,  $F(t_i)$  is the number of times the feature item appears in the text, and  $\vec{t}_i$  is its corresponding vector in the semantic space. After the feature vector is properly expressed, the text similarity between the student answer and the reference answer can be detected. Here, the vector cosine method is used for detection, and its expression is shown in formula (8).

$$SIM = \frac{\sum_{t=1}^N \omega_{it} \cdot \omega_{kt}}{\sqrt{\left(\sum_{t=1}^N \omega_{it}^2\right) \cdot \left(\sum_{t=1}^N \omega_{kt}^2\right)}} \quad (8)$$

In formula (8),  $SIM$  refers to the text similarity of student answers and reference answers.  $\omega_{it}$  and  $\omega_{kt}$  respectively represent the weight of student answers and reference answers in the  $t$  feature item.  $N$  is the total number of feature items. It can be seen that the smaller  $SIM$ , the smaller the text similarity, and vice versa. Finally, after the keyword matching degree and text similarity are obtained, the subjective questions can be scored according to their respective weights. The scoring calculation process is shown in formula (9).

$$S = (A \times S_k + B \times SIM) \times S_t, \quad A + B = 1 \quad (9)$$

In formula (9),  $S$  is the final score of students,  $S_k$  refers to the score of students in keyword matching,  $A$  and  $B$  are the weights of keyword matching and text matching respectively, and  $S_t$  is the total score of the topic.

C. Subjective Item Scoring Algorithm Based on Nonlinear Classifier KNN

As a compulsory course for most majors, College English courses attract a large number of students every year, which leads to a large number of examination papers on the WEB English teaching platform [24]. In order to further improve the efficiency of subjective question marking, KNN algorithm is introduced into the test paper scoring algorithm. The data scored by word matching and text similarity algorithm is used as the training set to train the KNN algorithm. The successfully trained KNN algorithm will be able to evaluate other test papers with high efficiency. The judgment principle of KNN Algorithm in subjective question scoring situation is shown in Fig. 5. For the red circular judgment object in the figure, KNN algorithm will calculate the samples of orange Pentagram and black triangle, that is, the distance between the training sample and the judgment object, take the first k distances with the shortest distance, and then analyze the category of the corresponding K samples. The category with the largest number of samples is considered as the category of the judgment object.

Although KNN algorithm has the advantages of fast operation and no need to retrain when adding new samples, when there is difference in the number of samples or uneven density distribution, it will also lead to great error in the output results [25]. Fig. 6(a) is a schematic diagram of the output error of the algorithm result caused by the error of the sample number. As shown in the figure, when the value of K is large, although the object to be determined is closer to Y, it may still be determined as X, because the number of X is much higher than Y. Fig. 6(b) is a schematic diagram of misjudgment caused by too large difference in sample density. It can be seen

that under this condition, the object to be judged is closer to Y, but the X distribution in a K finger is too dense, resulting in the number of X greater than Y.

In view of this situation, the KNN local weight correction algorithm is used to improve. The principle of the algorithm is to give a lower weight to the samples with too many and too large density compared with other training samples within the range of K value. On the contrary, a higher weight is given to smooth out the error. To describe the correction algorithm, a weight correction parameter needs to be defined, and its expression is shown in equation (10).

$$\left\{ \begin{aligned} \omega(c) &= \frac{\log(\frac{AvgNum}{Num(c)} + \beta)}{\log(\beta + 1)} \\ \beta &= \left\lfloor \frac{MaxNum}{AvgNum} \right\rfloor \end{aligned} \right. \quad (10)$$

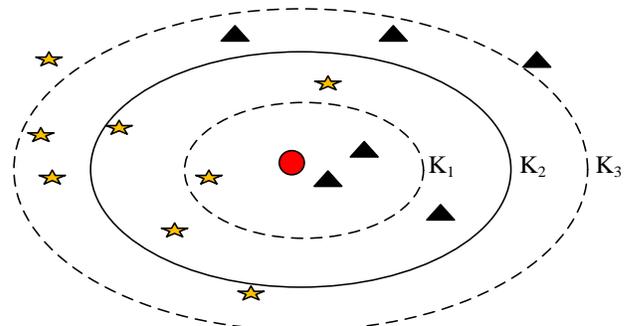


Fig. 5. Principle of KNN algorithm applied to test paper scoring.

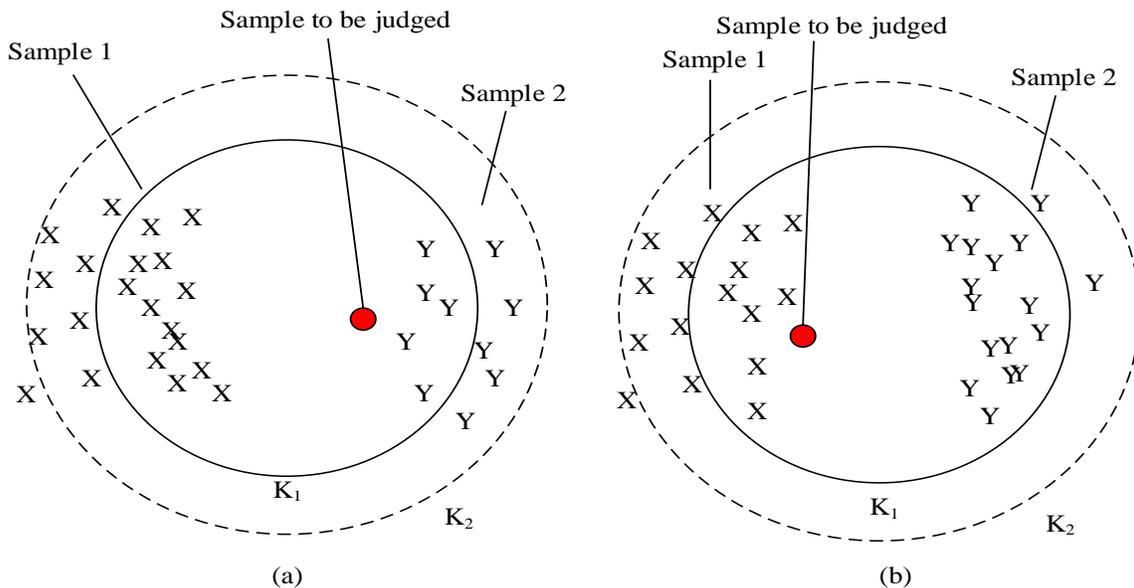


Fig. 6. Misjudgment principle of KNN Algorithm in subjective question scoring.

In equation (10),  $\omega(c)$  is the weight correction parameter of the object category,  $\beta$  is the adjustable parameter,  $MaxNum$  and  $avgNum$  are the maximum number of samples and the average number of samples of each category respectively, and  $Num(c)$  is the number of samples of the object category. The weight correction parameter can give different values according to the number and density of the actual training sets to smooth out this difference. The training samples with large differences in the number of samples can also enable KNN to output correct results. Finally, score one by one based on the weight correction parameters, and the expression of the final score is shown in equation (11).

$$S = \frac{\sum_{i=1}^k \omega(c) \cdot SIM}{k} \quad (11)$$

In equation (11),  $k$  is the nearest number,  $SIM$  represents the text similarity between the student's answer and the nearest sample, and  $S$  is the final score of the student's answer. Due to the existence of weight correction parameters, the weight of each type of sample is no longer unified as 1. Therefore, theoretically, the probability of outputting wrong results due to the difference in the number and density of samples will be greatly reduced.

#### IV PERFORMANCE ANALYSIS OF TEST PAPER SCORING ALGORITHM FOR WEB-BASED ENGLISH TEACHING PLATFORM

The performance analysis of the test paper scoring algorithm of the web network teaching platform mainly includes the judgment ability of the improved KNN algorithm, the differences between the scoring algorithm and manual scoring, and the scoring time. For KNN algorithm, the selection of K value has a great impact on its performance. Therefore, the algorithm is tested under different K values. The results are shown in Fig. 7.

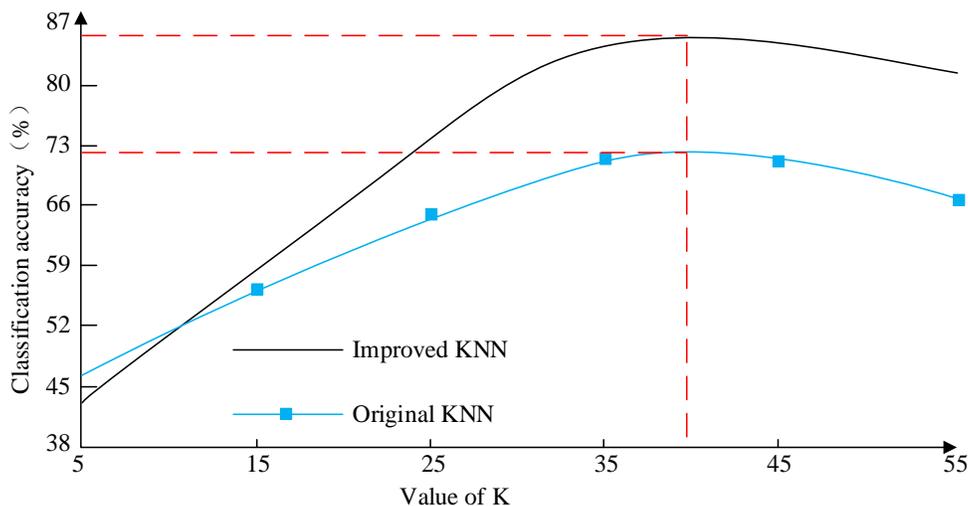


Fig. 7. Algorithm performance under different K values.

When the value of K is 39, the accuracy of both the improved KNN algorithm and the original KNN algorithm reaches the maximum, and then the accuracy of both algorithms begins to decline slowly. However, when K is 39, the accuracy of the improved KNN algorithm is 13% higher than that of the traditional algorithm, which shows that the weight correction parameters can significantly improve the accuracy of the KNN algorithm under the appropriate K value. Therefore, the value of K in this experiment is 39. The experiment was conducted on a WEB teaching platform based on Windows 10, which uses MySQL 5.1 database and Tomcat 6.0.33 server. In the process of correcting the actual test paper, different semantic topics may have an impact on the accuracy of the algorithm. Therefore, the common semantic topic data sets in six English tests are used to test the performance of the algorithm under different semantic topics. The performance is shown in Fig. 8.

Fig. 8(a) shows the test results of algorithm accuracy, Fig. 8(b) shows the test results of recall, and Fig. 8(c) shows the test results of F value. Under different semantic topics, the performance of the improved KNN algorithm and the original KNN algorithm shows obvious fluctuations. The improved KNN algorithm can achieve a recognition accuracy of 100% at most, while the lowest is only 34%. The highest recall rate is 0.90, and the lowest is only 0.50. However, compared with the two algorithms, the accuracy of the improved KNN algorithm is always higher than the original KNN algorithm, and the maximum difference between the two is 0.50. Except for the sixth semantic topic, the recall rate of the improved KNN algorithm is also higher than the original KNN algorithm. The results show that different semantic topics may have a significant impact on the performance of the algorithm. The web-based test paper scoring algorithm is constructed by imitating the mechanism of teacher manpower scoring. Therefore, comparing the scoring results with the teacher manpower scoring results can better evaluate its performance. The comparison results are shown in Fig. 9.

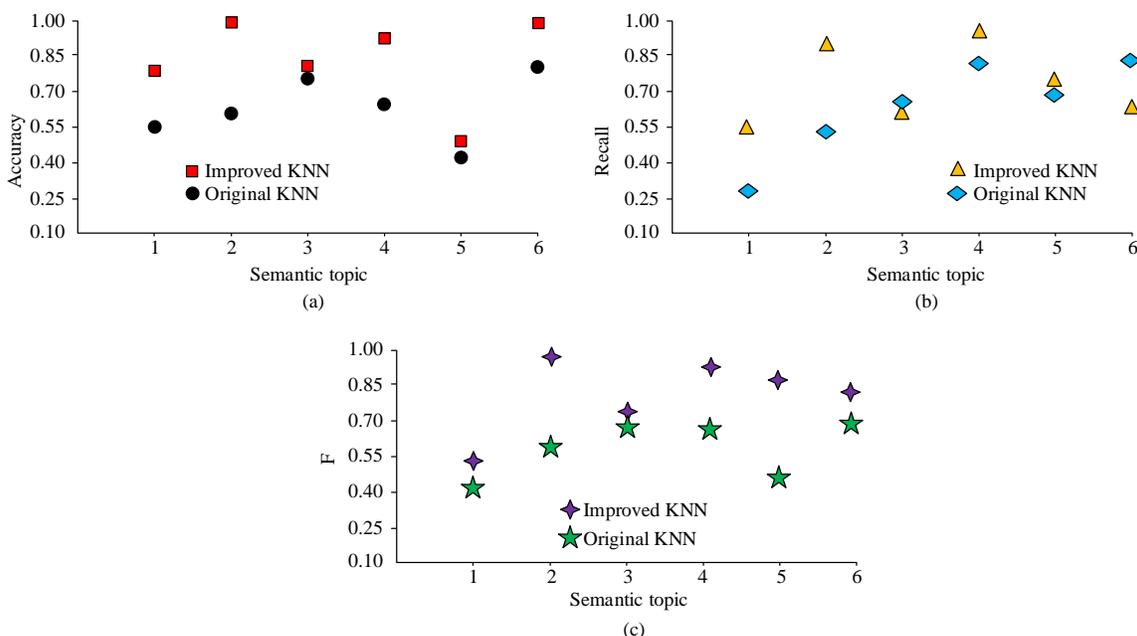


Fig. 8. Performance of the algorithm under different semantic topics.

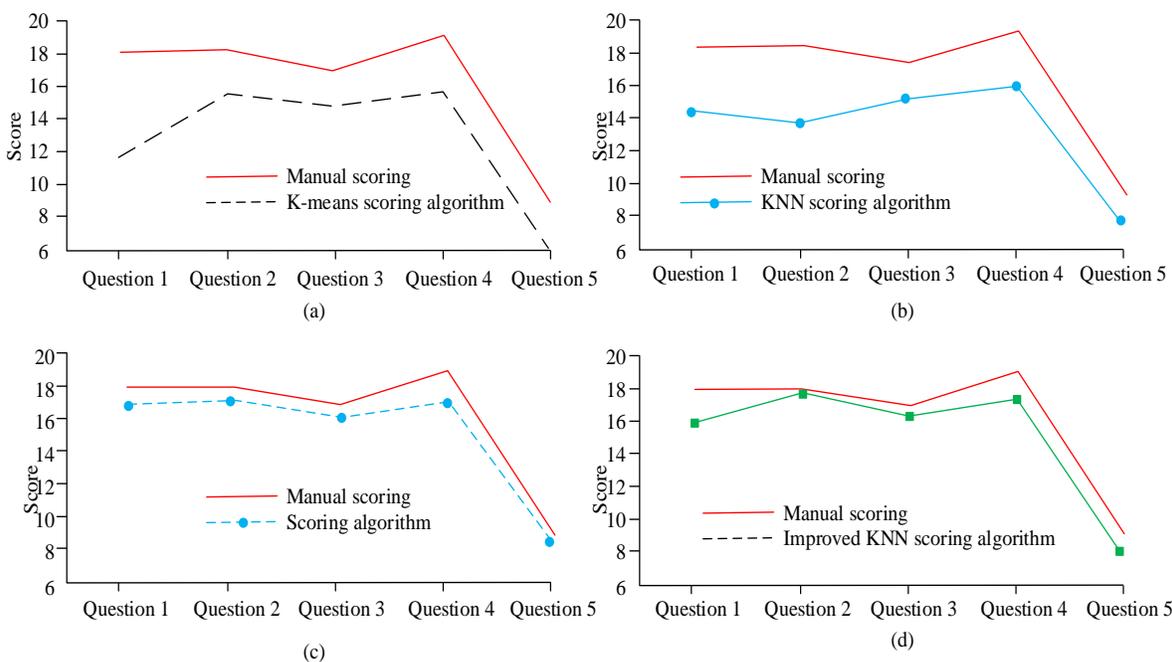


Fig. 9. Comparison of different algorithms and manpower scoring.

Fig. 9(a), Fig. 9(b), Fig. 9(c) and Fig. 9(d) are the comparison results of human scoring and K-means clustering algorithm, original KNN algorithm, test paper scoring algorithm without KNN and improved KNN algorithm respectively. Three machine learning algorithms are trained based on test paper scoring algorithm. It can be seen that the change trend of the scores of the four algorithms is basically consistent with the human scoring, which means that the four algorithms have successfully imitated the mechanism of teachers' human scoring to a certain extent, but the distance

between the broken line of K-means clustering algorithm and the original KNN algorithm and the broken line of human scoring is significantly greater than that of the test paper scoring algorithm and the improved KNN algorithm, which means that the test paper scoring algorithm and the improved KNN algorithm have a better effect on the imitation of human scoring. In order to further study the performance differences of several algorithms, the difference between them and the human score is described with pictures, as shown in Fig. 9.

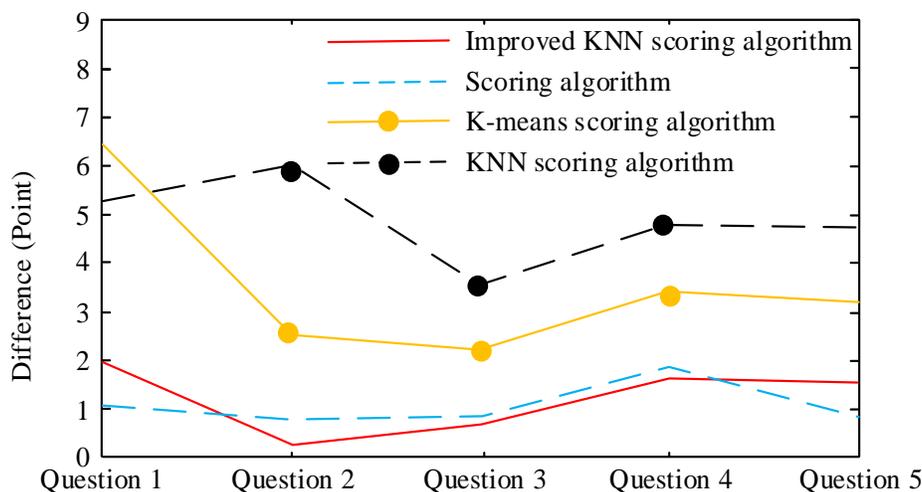


Fig. 10. Difference between different algorithms and manpower scores.

It can be seen from Fig. 10 that the difference between the scoring algorithm and the improved KNN algorithm and the manpower score is very low, ranging from 0 to 2 points, and the difference between the two is also small, less than 1 point. The difference between the other two algorithms is significantly greater, and the difference with the manpower score fluctuates between 2 and 6.5 points, which indicates that the scoring algorithm has a good imitation effect on the manpower score, while the improved KNN algorithm has a good learning effect on the scoring algorithm, and the learning effect of the original KNN and K-means clustering algorithm is inferior to the improved KNN algorithm. In addition to teachers, students often evaluate the fairness and accuracy of the automatic scoring algorithm. With the student feedback system of the Web English teaching platform, we studied and collected the proportion of misjudgments reported by students in multiple test papers through the platform under several scoring algorithms, and evaluated the performance of the algorithm from this angle. The results are shown in Fig. 11.

By comparing several algorithms, it is found that the proportion of students' reported misjudgment under the original KNN algorithm is the highest in each test paper, the highest is 12.5%, and the lowest is 7.8%. The student report misjudgment ratio of the scoring algorithm and the improved KNN algorithm is always lower than that of the original KNN algorithm, of which the highest is 10.8% and the lowest is 7.3%. According to the data statistics of the web platform, the average student report misjudgment ratio of teachers' manual correction is 5.7%. The algorithm is very close to this standard, which means that the evaluation accuracy of the algorithm is also high from the perspective of the evaluated students. Finally, the time consumed by different algorithms for the same test set is studied and counted. Since the original intention of the scoring algorithm is to improve the efficiency of the Web English teaching platform, the algorithm time is an important evaluation item. The results of time-consuming evaluation are shown in Table I.

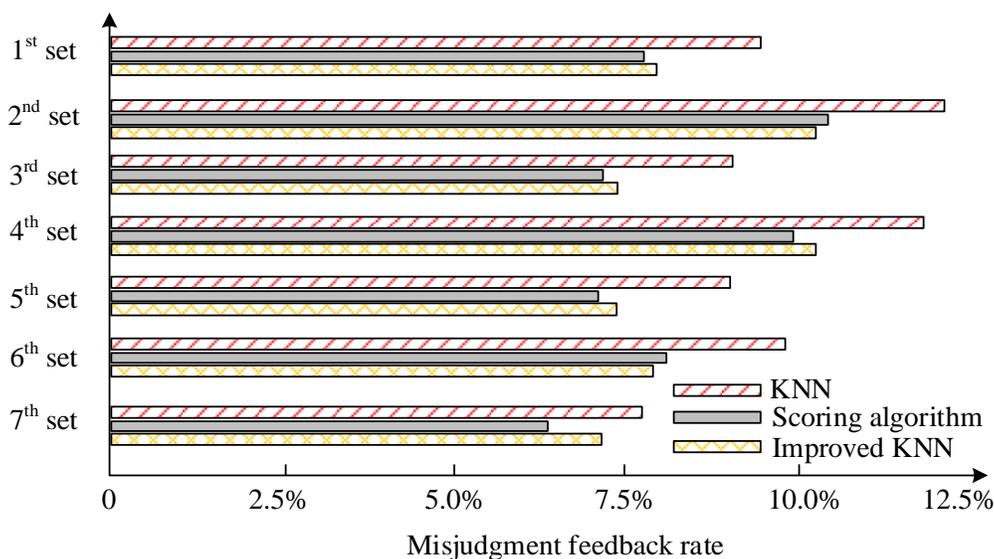


Fig. 11. Proportion of students' report misjudgment under different algorithm.

TABLE I. ALGORITHM TIME-CONSUMING DETECTION

		Scoring algorithm	Improved KNN	KNN
Time consuming (s)	Set 1	177	124	137
	Set 2	130	105	100
	Set 3	212	108	114
	Set 4	97	63	59
	Set 5	133	93	91

Table I describes the time-consuming of scoring three algorithms in five different sets of test papers. The time-consuming of the original KNN algorithm and the improved KNN algorithm is always less than that of the scoring algorithm for each set of test papers. The difference between the time-consuming of the improved KNN algorithm and the scoring algorithm is up to 104 seconds, indicating that the machine learning algorithm is stable in judging speed than the scoring algorithm without machine learning. Comparing the original KNN algorithm with the improved KNN algorithm, it is found that the time-consuming of the two algorithms is relatively close, and they have their own advantages and disadvantages in different test papers, which shows that the improved KNN algorithm is similar to the original KNN Algorithm in terms of calculation speed.

## V DISCUSSION

The English grading algorithm based on KNN and text similarity is constructed. The algorithm is divided into two parts: objective question scoring and subjective question scoring. Due to the inconsistency between students' answers and reference answers, it has been difficult to use automatic algorithms to completely replace teachers' manual scoring in the subjective scoring of English test papers. The algorithm's ability to judge the text similarity of different semantic topics has been tested. The results show that the proposed algorithm can achieve the highest recognition accuracy of 100%, and the highest recall rate is 0.90. Even for the performance of the method itself, the improved KNN structure in the algorithm is obviously superior to the ordinary KNN model. In the test of actual English test paper data, the algorithm is used to compare with the teacher's manual grading. Compared with other similar algorithms, the score given by the proposed algorithm is significantly closer to the score of the teacher's manual score, and the maximum difference between the scores is no more than two points. Further research on the misjudgment rate reported by students, shows that the proposed algorithm has the lowest misjudgment rate, which is the closest to the misjudgment rate of teachers' manual grading. In the current subjective question scoring algorithm applied in the online learning platform, the collected data of false judgment rate is often more than 13%. Therefore, the research believes that the proposed algorithm has higher scoring performance in comparison, and can be applied to the English test paper scoring on the WEB online learning platform.

## VI CONCLUSION

The WEB-based College English course network teaching cooperation platform has broadened the channels of College

English teaching, so that students and teachers can carry out English teaching activities more conveniently and efficiently. In the online examination of the WEB College English teaching platform, the scores of test papers, especially the subjective questions, are often scored by teachers' manpower, which is no different from the efficiency of traditional offline teaching. Therefore, this research designs a test paper scoring algorithm based on the College English teaching platform combined with the improved KNN algorithm. The performance test results show that the algorithm performs well in the similarity between the scores and the scores of teachers' manpower, The lowest score difference between it and the teacher manpower score is only 0.4 points, and the highest is only 2 points. In addition, the algorithm has outstanding performance in the classification accuracy of different semantic topics. The accuracy of some semantic topics reaches 100%, and the accuracy of all semantic topics is higher than the traditional KNN algorithm. In terms of the time-consuming of the algorithm, the minimum time-consuming of the algorithm in the experiment is only 63 seconds, which is significantly faster than the human scoring speed. According to the test results, the algorithm can correct the objective and subjective questions of the online English teaching test paper with the accuracy close to that of human marking. Its application can effectively reduce the workload of teachers and improve efficiency. At the same time, the algorithm has the potential to be applied to other subjects. The imperfection of this study lies in the calculation speed. The improved KNN algorithm is not much different from the traditional algorithm. Therefore, on the basis of this study, how to improve the speed is the next research direction.

## REFERENCES

- [1] Q. Yuan, "Network education recommendation and teaching resource sharing based on improved neural network", *Journal of Intelligent and Fuzzy Systems*, vol. 39(4), pp. 5511-5520, 2020.
- [2] H. Zaini, A. Oni, A. Hadi, et al. "Covid-19 and Islamic Education in School: Searching for Alternative Learning Media", *Webology*, vol. 18(1), pp. 154-165, 2021.
- [3] J. Das, S. Majumder, P. Gupta, et al. "Collaborative Recommendations using Hierarchical Clustering based on K-d Trees and Quadrees", *International Journal of Uncertainty Fuzziness and Knowledge-Based Systems*, vol. 27(4), pp. 637-668, 2019.
- [4] E. Tsalera, A. Papadakis, M. Samarakou, "Monitoring, profiling and classification of urban environmental noise using sound characteristics and the KNN algorithm", *Energy Reports*, 2020, vol. 6(6), pp. 223-230, 2020.
- [5] Poultakis, S., Papadakis, S., Kalogiannakis, M., & Psycharis, S. (2021). The management of Digital Learning Objects of Natural Sciences and Digital Experiment Simulation Tools by teachers. *Advances in Mobile Learning Educational Research*, 1(2), 58-71.
- [6] Papadakis, S., Kalogiannakis, M., Sifaki, E., & Vidakis, N. (2017). Access moodle using smart mobile phones. A case study in a Greek University. In *Interactivity, Game Creation, Design, Learning, and Innovation* (pp. 376-385). Springer, Cham.
- [7] Panagiotakopoulos, C., Daloukas, V., & Panagiotakopoulos, T. (2022). Strengthening the coding skills of teachers in a low dropout Python MOOC. *Advances in Mobile Learning Educational Research*, 2(1), 187-200.
- [8] A M. Christianson, "Using Socratic Online Polls for Active Learning in the Remote Classroom", *Journal of Chemical Education*, vol. 97(9), pp. 2701-2705, 2020.
- [9] Karakose, T., Ozdemir, T. Y., Papadakis, S., Yirci, R., Ozkayran, S. E., & Polat, H. (2022). Investigating the relationships between COVID-19

- quality of life, loneliness, happiness, and internet addiction among K-12 teachers and school administrators—a structural equation modeling approach. *International Journal of Environmental Research and Public Health*, 19(3), 1052.
- [10] <https://doi.org/10.25082/AMLER.2022.01.003> Lavidas, K., Apostolou, Z., & Papadakis, S. (2022). Challenges and opportunities of mathematics in digital times: Preschool teachers' views. *Education Sciences*, 12(7), 459.
- [11] Katsaris, I., & Vidakis, N. (2021). Adaptive e-learning systems through learning styles: A review of the literature. *Advances in Mobile Learning Educational Research*, 1(2), 124-145.
- [12] Z. A. Zardari, J. He, M. S. Pathan, et al. "Detection and Prevention of Jellyfish Attacks Using kNN Algorithm and Trusted Routing Scheme in MANET", *International Journal of Network Security*, vol. 23(1), pp. 77-87, 2021.
- [13] B. Wang, X. Gan, X. Liu, et al. "A Novel Weighted KNN Algorithm Based on RSS Similarity and Position Distance for Wi-Fi Fingerprint Positioning", *IEEE Access*, vol. 8(1), pp. 30591-30602, 2020.
- [14] Z. Lei, L. Zhu, Y. Fang, et al. "Anomaly detection of bridge health monitoring data based on KNN algorithm", *Journal of Intelligent and Fuzzy Systems*, 2020, vol. 39(4), pp. 5243-5252, 2020.
- [15] S. C. Pang, et al. "A Text Similarity Measurement Based on Semantic Fingerprint of Characteristic Phrases", *Chinese Journal of Electronics*, 2020, vol. 29(02), pp. 39-47.
- [16] S. Yang, Y. Tang, "News Topic Detection Based on Capsule Semantic Graph", *Big Data Mining and Analytics*, vol. 5(2), pp. 98-109, 2022.
- [17] R. Franclinton, O. Karnalim, M. Ayub, "A Scalable Code Similarity Detection with Online Architecture and Focused Comparison for Maintaining Academic Integrity in Programming", *International Journal of Online and Biomedical Engineering (iJOE)*, 2020, vol. 16(10), pp. 40-52, 2020.
- [18] Артур Гудманиян, Сергей ванович Сидоренко. "A Web Resource Map for A University Course in The History of the English Language", *Information Technologies and Learning Tools*, vol. 76(2), pp. 126-136, 2020.
- [19] A. Miltiadous, D. L. Callahan, M. Schultz, "Exploring Engagement as a Predictor of Success in the Transition to Online Learning in First Year Chemistry", *Journal of Chemical Education*, 2020, vol. 97(9), pp. 2494-2501, 2020.
- [20] M. Liu, N. Noordin, L. Ismail, et al. "The Role of Smartphone Applications as English Learning Tool among Chinese University Students", *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, 2021, vol. 12(14), pp. 385-398, 2021.
- [21] I. U. Khan, M. Ullah, K. Iqbal, et al. "Perceptions of Efl Learners Regarding Effectiveness of Twitter in English Language Learning Proficiency at University Level", *Humanities & Social Sciences Reviews*, vol. 9(3), pp. 1355-1365, 2021.
- [22] P. Mehta, S. Chandra, "Robustness and Predictive Performance of Homogeneous Ensemble Feature Selection in Text Classification", *International Journal of Information Retrieval Research*, vol. 11(1), pp. 75-89, 2021.
- [23] Y. Wang, "Similarity detection of English text and teaching evaluation based on improved TCUSS clustering algorithm", *Journal of Intelligent and Fuzzy Systems*, vol. 40(4), pp. 1-11, 2020.
- [24] L. L. Jassim, "Methods for Learning English Vocabulary Employed by Iraqi EFL Learners at University of Technology", *Arab World English Journal*, vol. 2(2), pp. 314-328, 2021.
- [25] Y. Ma, X. Zhao, "POD: a Parallel Outlier Detection Algorithm Using Weighted kNN", *IEEE Access*, vol. 9(1), pp. 81765 – 81777, 2021.

# Predictions of Cybersecurity Experts on Future Cyber-Attacks and Related Cybersecurity Measures

Ahmad Mtair AL-Hawamleh  
Department of Electronic Training,  
Institute of Public Administration, Riyadh, Saudi Arabia

**Abstract**—The Internet interconnections' exponential growth has resulted in an increase in cyber-attack occurrences with mostly devastating consequences. Malware is a common tool for performing these attacks in cyberspace. The malefactors would either exploit the present weaknesses or employ the distinctive characteristics of the developing technologies. The cybersecurity community should increase their knowledge on the types and arsenals of cyber-attacks, and security measures against cyber-attacks should be in place as well. Also, advanced and effective malware defense mechanisms should be established. Hence, this study reviews cyber-attack types, measures and security precautions, and professional extrapolations on cyber-attacks future and the associated security measures. Semi-structured interviews were performed, involving five IT managers and nine Cybersecurity Consultants, to obtain the data. The study findings demonstrate prevention as key for data breach risk prevention. Knowledge of common attack methods and the use of cybersecurity software can facilitate individuals and organizations in thwarting hackers and in preserving their data privacy. Two-factor authorization by consumers and new back-end security protocols and security methods, including Artificial intelligence (AI) application, will encumber hacking attempts.

**Keywords**—Cybersecurity; cybercriminals; cyber-attacks; cyber-security techniques; security precautions

## I. INTRODUCTION

The last 20 years have seen the emergence of the Internet as key to global communication, making the Internet today an important part of the life of people globally. Worldwide, [1] reported that there were more than 3 billion internet users, and such number of users has been greatly factored by the ubiquity and low cost of the Internet. Through the immense global network made possible by the Internet, billions of dollars have been generated each year, adding to the global economy [2, 3]. Today, interactions and many other activities of various types are being carried out in cyberspace [3,4], like financial transactions between business parties and casual communication between individuals [5,6]. The cyberspace controls, manages and exploits vital and sensitive infrastructures and systems, but the cyberspace can also be made up of such infrastructures and systems [7].

The cyberspace is where varied aspects of citizens' lives are interconnected, and so, as indicated by [8], the conditions of the cyberspace will directly affect the lives of the citizens [8]. Today, cyberspace has presented new security challenges to governments. As reported by [9], cyber-attacks and their effects have been the subject of concern among analysts since the last decade. Cyber-attacks undeniably lead to damages, physical and/or economic, and there have been cases where the damages were so widespread and severe, as can be exemplified

by virus attacks on stock market of a country, resulting in a crash or a loss of colossal amount of money, or virus attacks on a power plant system leading to a massive explosion loss of lives and properties, just to name a few [10, 11].

The advent of the Internet and the current digital transformation have made cybersecurity a serious matter to experts and all other involved parties. Today, people are increasingly dependent on computer systems, the Internet, and wireless networks (e.g., Bluetooth and WiFi) [12], as can be observed by the increasingly common dependence of people towards smart devices like smartphones and the Internet of Things. In other words, people today are more at risk of being victims to cyber-attacks. Meanwhile, as reported by [13], the discovery of malicious activities on the network has been too common. In general, detection of intrusion has been reactive in nature, and the reaction is only to certain patterns or observed anomalies. However, scholars including [3] and [14] mentioned the need to employ a proactive approach instead, that is, to react to these intrusions before they manage to cause any harm. Somehow, it appears that work and advancement in predictions, measures and security precautions concerning cyber security remain obscure, but, recently, the efforts seem to be gradually gaining momentum [3, 14, 15].

Cyber-attacks are becoming more complex and severe as time passes [16]. Currently, only little is understood pertaining to the different types of cyber-attacks, how these attacks spread, and the current security precautions against them. As such, many organizations/countries have fallen victim to these attacks. Meanwhile, security measure development needs comprehensive understanding of these attacks, and so, a complete listing and classification of cyber-attacks is a crucial element of cyber security initiatives. The present study seeks to describe the different kinds of cyber-attacks, measures and security defenses against them, while also predicting future cyber-attacks. Some security measures are proposed, to facilitate programmers in their development of security devices and mechanisms according to the mode of attack.

## II. RELATED WORKS

### A. Cybersecurity

Cybersecurity entails a bulk of practices, technologies, and processes specifically developed to safeguard the data, networks, programs, and devices of people and enterprises from attacks [3], [14]–[18]. Meanwhile, financial, corporate, government, military, and medical organizations generally gather, process, and store data in substantial amounts on computers and other devices. Some of these gathered data

can be sensitive data like personal data, financial data, and intellectual property, and so, access to such data requires authorization because negative consequences could result when these data are accessed by unauthorized parties [15, 19, 20]. Therefore, for these organizations, cybersecurity is a crucial matter [19]–[22].

In business transactions, sensitive data is transferred to other devices across networks. Appositely, cybersecurity encompasses the discipline for safeguarding the information and the applied systems in its process or storage [23]–[26]. In this regard, firms that are responsible to protect the information on financial records, health, and national security are obliged to take measures to safeguard their sensitive business and personnel information, especially now that cyber-attacks have increased in terms of volume and superiority [23]–[26].

Through a solid cybersecurity strategy, an appropriate security mechanism can be achieved, and this mechanism could effectively deter malicious attacks that generally would attempt to access, modify, erase, extinguish or extort the systems and sensitive data of persons or organizations [27]–[29]. Additionally, cybersecurity could deter attacks that could incapacitate or mess up the operations of a device or system [27]–[29].

### B. Cybercriminals

Cybercrime relates to criminal activity involving a computer, networked device or a network [30]–[34]. In general, cybercrimes are executed by cybercriminals for the purpose of making personal gain [30]–[34]. However, there are cybercrimes performed for the purpose of damaging or disabling computers. There are also those who utilize computers or networks for the purpose of disseminating malware, and also for dispersing prohibited information, images or other materials [30]–[34].

Cybercrime was defined by the Council of Europe Convention on Cybercrime as a vast gamut of malicious activities, and among these activities are unlawful data interception, copyright infringements, and system intrusions, that impair the integrity and availability of the network [35, 36]. The USA is a signatory to this council [35, 36].

The internet connectivity is a common availability today as it is a requirement to various daily undertakings. However, such availability has caused cybercrime activities to thrive as the culprit could commit the crime without having to be physically present [37]. Fraud, money laundering, cyberbullying and cyber stalking are among the examples of commonly committed cybercrimes, and these crimes are further facilitated by the speed and convenience of the internet, and the anonymity and borderless reachability that the Internet is offering [23, 28, 36, 37].

Cybercrimes may be executed by persons or groups with fairly little technical skills. Equally, the crimes may be committed by extremely structured global criminal groups involving skilled developers and other experts. Also, it is common to see cybercriminals operating in countries that have no or weak cybercrime laws so that they could not be easily detected or prosecuted [38, 39].

### C. Cyber Attacks

A cyber-attack is an intentional and malicious effort made by a person or an organization to break the information system of others [40]. The attack is usually economically driven, but there are also attacks that involve data or information stealing, modifying or destruction. In other words, among the goals of attack include to break the system, or to steal, modify or destroy the data or information of others [3, 15, 39, 40].

Cyber-attacks are more and more common these days. Furthermore, the Cisco Annual Cybersecurity Report [41] has relevantly indicated that the advent of network-based ransomware worms has allowed attackers to launch campaigns without the need for human involvement. Also, security events nowadays have become more intricate and more copious [41]. Moreover, businesses today face cyber-attacks on a daily basis; it was mentioned by the then CEO of Cisco (Mr. Chambers) that businesses can be classed into two groups; One comprises those that have been hacked, and the other comprises those that are still unaware of the fact that they have been hacked [42].

Cyber-attacks generally occur in six forms namely: Malware, Phishing, Denial of Service (DoS), Man-in-the-Middle (MitM), Password Spraying, and Cross-site Scripting (XSS) [40]–[49]. Each of these attack types is described as follows:

1) *Malware*: Malware encompasses malicious code or malicious software, and it is essentially a program that is covertly implanted inside a system with the purpose of disrupting the data so that the data would lose their integrity, confidentiality, or accessibility [50]. Malware is regarded as a major external threat to systems as it can affect the systems' operation [40, 43, 50]. Malware can cause widespread damage and disruption, and significant efforts would be required to fix this malware problem. Malware comes in various forms including Trojans, virus, worms, spyware and ransomware [40, 43, 50]. The details are as follows:

- Trojans: This type of malware is also known as Trojan horse, and encompasses a seemingly legitimate and safe file, program, or piece of code (but indeed a malware) [43, 51]. Usually, Trojans are bundled and transported within an authentic software, and are created for spying on or for stealing data from victims. Trojans display themselves as genuine files, and so, victims would be misled to click, open, or install these Trojans (without knowing). Upon installation, many Trojans will download other malware to spy on the victim or cause other types of harm.
- Viruses: Viruses generally will attach themselves to the order of initialization, and these viruses would replicate themselves to infect other codes within the computer system [40, 51, 52]. They also could attach themselves to executable code or link themselves with a file through forming a virus file with a similar name but with an extension [51, 52]. This file is a decoy that transports the virus [40, 51, 52].
- Worms: Worms encompass self-contained programs spreading across networks and computers [43, 51, 53]. Frequently installed via email attachments, worms would dispatch a copy of themselves to all contacts

in the affected computer email list [53]. Usually, worms are used by perpetrators to overload an email server and generate a denial-of-service attack. However, worms don't attack the host like viruses do [40,53].

- **Spyware:** This type of malware entails a program that attackers use to gather information relating to users, their systems or browsing routines, sending the data to a remote user [54]. The obtained information can be used by the attacker to blackmail the user. Also, the attacker could download and install other malicious programs from the web [40, 43, 51, 54].
- **Ransomware:** Ransomware is a very common attack method with the ability to inhibit or restrict the access of users to their system [55]. It also may instruct users to pay a certain amount of ransom using online payment methods [56], which generally would involve the use of virtual currencies like bitcoins, before they could re-access their system or data. Ransomware gets into computer networks; through the use of public-key encryption, ransomware encrypts the files, and this encryption key remains with the server of the cybercriminal [56]. Encryption is used by cyber criminals to detain the data, and the data owner has to pay a certain amount of ransom to get the private key [40, 43, 55, 56].

2) *Phishing:* Phishing is an activity of transmitting deceitful communications through seemingly reputable emails [45,47]. It is common to see these emails demonstrating legitimacy but they actually link the receiver to a malicious script or file [45]. Through this script or file, the attackers could gain access to the device of the victim and gain control over it. Consequently, the attacker could also insert malicious scripts/files, and extract sensitive data like user information, financial data, and so forth [45,46]. Essentially, phishing is done to steal confidential data such as the victim's login information and credit card details.

3) *Denial of service (DoS) and distributed denial-of-service (DDoS):* DOS attacks involve flooding the systems, servers, and/or networks with traffic for the purpose of overloading the resources and bandwidth, resulting in failure of the system in meeting valid requests [48]. DoS attacks can be simultaneously executed by various computers at one specific time and this is called Distributed Denial-of-Service (DDoS) attack [34, 48]. Dealing with DDoS attacks can be very challenging because attackers can come from various IP addresses globally, making it very difficult for network administrators to determine the attack source [30, 48].

4) *Man-in-the-middle (MitM):* Man-in-the-Middle (MitM) attack which is also called eavesdropping attack, involves hijacks by an attacker during a session between a trusted client and network server [47]. During the attack, the attacker's computer switches the IP address of a trusted client while the session is resumed by the server as the server thinks that it is still in communication with the client, not knowing that the client has been replaced with the attacker's computer [36, 47]. As an illustration: a client is in connection with a server when the computer of the attacker gains control over the client. The computer of the attacker then disconnects the client from the

server. This is followed by the replacement of the client's IP address with that of the attacker's computer. The sequence numbers of the client are spoofed. The communication between the network server and the client resumes but the server does not know that it is no longer communicating with the client, but with the attacker's computer instead.

5) *Brute-force and password spraying:* Brute-force attacks generally involve attacks on a single account, whereby the attacker would test various passwords in the attempt of gaining access to an account [44]. This leads to recurrent failed logins [44, 57]. However, in general, modern cybersecurity protocols are able to identify such activity and will lock out an account following several failed login attempts within a short period of time [57].

However, the use of password spraying by attackers can overturn the standard protocols of cybersecurity [58]. Hence, the attacker would try to log on to several user accounts with the use of various passwords that are commonly used. Using a single password on several accounts before using another password on the exact accounts would prevent the standard lockout protocols from being activated. This way, the attacker could continue trying out more and more different passwords on the target account [44, 57, 58].

Owing to the failure of many users in adhering to best practices of password usage, the method of password spraying attacks are often successful. As reported in 2019, recognizable number arrangements like "12345", typical names of females like Jennifer, and the word "password" are among the most commonly used passwords among users [59, 60]. These, and other reported 200 easily guessed passwords have contributed to data breaches [59, 60]. Hence, attackers targeting a reasonably large number of usernames and utilizing a sufficiently large array of common passwords are likely to succeed in gaining access to some accounts [40, 43, 44, 59, 60].

6) *Cross-site scripting (XSS):* Cross-site scripting or XSS refers to a weakness of web security, allowing attackers to compromise the interactions of users with weak applications [49, 61]. The weakness of the user's system allows attackers to evade the exact origin policy that distinguishes websites [62]. The attacker could disguise as the user and perform any actions of the user and access all data of the user. Hence, the attacker may have complete functionality control over the application belonging to the user if the user has privileged access within the application [49, 61].

#### D. Cybersecurity Tools and Techniques

Today, the number of illegal attempts to gain access to private data has increased. These attempts are generally for stealing the data or for forcing users into information blackmailing. Such a situation has increased the importance of cybersecurity [3, 14, 15, 17, 18]. There are various methods being used in achieving cybersecurity. Among them include anti-virus, firewall, authentication, encryption and digital signatures. The details of each are as follows:

- **Anti-Virus:** A computer virus is generally an unwanted short program that prompts undesirable commands without user consent. An antivirus generally performs two tasks [15, 63]. The first task is to prevent the

installation of a virus in a system, and the second task is to scan the system to find out if there are viruses existing in a system [15, 63]. Most viruses are created to attack Windows operating systems because most users prefer to use Windows as their computing platform. However, there are also viruses created to attack Apple and Linux operating systems [63, 64].

- **Firewall:** Firewalls are created to provide an effective deterrent towards hackers' attempts to illegally access a computer upon its connection to the internet or to other network connections [65, 66]. Most operating systems are equipped with a firewall and the firewall is turned on by default. In addition to the default firewall, users could also install commercial firewalls if the default firewall does not provide sufficient protection or if it interferes with the user's legitimate network activities [65, 66].
- **Authentication:** Authentication is regarded as the basic method of cybersecurity. It provides a verification to user identity according to the records saved in the system's security domain [67]. Password technology is the most common security control method, but there are also other methods including SIM cards that are inserted into the cell phone of the user [67]–[69]. A SIM card has unique ID numbers, and during identification of certain cell phones, these numbers are transmitted over a secure communication line. During the process of authentication of a message, there may be eavesdropping attempts made by unauthorized parties, and it may be very difficult to counter these attempts. The transmission of password through an insecure medium may cause the password to be intercepted by fraudulent individuals disguising as the original user. Encryption can be used in dealing with this problem [67]–[69].
- **Encryption:** Encryption makes data incomprehensible and the right key is required to unlock it [70]. Resolving an encryption requires a resolution of intricate mathematical problems (e.g., factoring large primes) which is highly resource and time consuming [70, 71]. In the encoding and decoding of a message, a similar key is used in symmetric encryption, with security level similar to that of the key. Possible security risks are included alongside the key distribution. For asymmetric encryption, the public key is used in message encryption, while the private key is used in message decryption. For key distribution in today's security protocols, most employ asymmetric encryption [71].
- **Digital signatures:** Digital signatures can be formed from similar mathematical algorithms applied in asymmetric encryption [72]. Utilizing some information encoded with it, the user could perform a test to see if the key that he is in possession of, is private. Similar decryption can be obtained by the user through the attainment of a public key that will verify his (user) credentials [72, 73]. In general, this process is identical to that of public key encryption, and it operates based on the assumption that the authorized user is the only party in possession of private key [72, 73].

### III. METHODOLOGY

The study data were obtained by way of semi-structured interviews. The interviews took 45 minutes, and the interviewees comprised five IT managers and nine cybersecurity Consultants. The interviewees were reached through email and they took part in the interviews willingly. They were asked about the types of cyber-attacks and the measures and security precautions to be taken to counter these attacks. The interviewees were also asked to make predictions concerning the future of cyber-attacks and relevant security measures. All interviews would be stripped off of the identifying details. The authors would discuss the extracted data until an agreement was reached.

### IV. FINDINGS AND DISCUSSIONS

The conducted interviews resulted in the understanding of cyber-attacks in terms of types and weapons. Also, the interviews provide the researcher with understanding of the measures and security precautions to counter cyber-attacks. Additionally, the interviewees made predictions concerning the future of cyber-attacks and the associated security measures. Accordingly, all interviews would be stripped off of identifying details. Any disagreements or concerns about the extracted data were discussed among the authors until a consensus was reached.

#### A. Types and Weapons of Cyber Attacks

There is an increasing number of people working from a distance or online since the year 2019 [74]. Consequently, as reported by the FBI, the number of cyber-attacks have increased by fourfold towards online activities [74]. Also, studies have shown an increase in the impact of successful cyber-attacks on organizations and their users each year [57, 74, 75].

As reported, about 197 million records became exposed in 2017 because of data breaches, and the number of exposed records rose to 37 billion in 2020 albeit the decrease in the overall number of data breaches [76, 77]. In countries such as the USA, IBM reported that organizations have to incur increasing cost when facing these attacks. As reported, the average cost of a data breach has risen to \$8.64M in 2020 from \$7.91M in 2018 [76, 77].

A lot of times, the success of cybercriminals in breaching organizations, also depending on the methods used, is facilitated or made possible by human error [3, 15, 17, 18]. For instance, the use of phishing may not be successful if the target victim did not click on the link provided. Also, some types of cyber-attacks exploit the gaps in user's efforts in data security, allowing these cybercriminals to gain access to sensitive data [3, 15, 17, 18].

In the interviews, the participants were asked on the types and weapons of cyber-attacks. One participant stated the following:

*“Up to now, we are still facing some of the most common cyber-attacks, for example, Password Spraying Attacks, Ransomware, Denial of Service, and Malware Attacks too. Fortunately, there was no serious damage because we had been prepared – we invested a lot to build strategies against cyber-attacks. Still, we are not saying that we were unharmed at*

all. The damage was there, of course. As precautions, we are consistent with our Advanced Persistent Threats (APT) audits. Also, we would always check our risky points.”

Concerning the topic of ransomware, another interview participant stated the following:

*“It appears that ransomware is not just a security incident. It has changed. Also, it is obvious that the present-day cybercriminals want to breach our data, and for that, they are teaming up with organized cybercrime groups to steal the data. Then, they encrypt on corporate servers. As a company, our focus is to regain our data. At the same time, we worry about which public is sharing the data. Cybercriminals use ransomware when they are under extreme pressure, and their targeted victim could be any party – individual, company, or government.”*

The interviewees were asked about Phishing techniques. The response of one interviewee is as follows:

*“Phishing techniques involve emailing – the attacker would email thousands of deceitful messages to target victims. For instance, the victims would receive an email on receiving a handsome amount of money. Out of thousands of emails sent, certain fractions of target victims would fall for the scam. To increase the success rate, the attackers would use certain techniques, like mimicking the actual emails from a spoofed organization. For instance, PayPal. So, they would use similar phrasing, typefaces, logos, and signatures as PayPal’s. So, the messages will look legitimate that the victim is less aware that they are being scammed or attacked.”*

One more interviewee offered an opinion regarding phishing, by stating the following: *“It is an increasingly common cyber threat today. As you can see, it is common to receive emails from what seems like a reputable source when it is really not the case. These criminals - they just want to steal sensitive data like your login and credit card information. They also want to install malware on your computer.”*

The interviewees were asked on the subject of Distributed Denial of Service techniques used by cybercriminals. An interviewee responded by saying:

*“Usually, a DDoS attack is done by a group of malware-infected host machines. The attacker controls these machines. We call these attacks the Distributed Denial of Service. This is because the attacks prevent the affected site from providing the user with the service it is supposed to provide, and therefore, the user cannot gain access to the site. The attack causes the victimized site to become flooded with illegal requests, and since the site has to answer to each request, its resources become all used up that it cannot serve users. A shutdown may happen to the site.”*

Research has confirmed that the attacks are essentially actions aimed at impairing a system or disturbing the normal operations of a system through the exploitation of the system’s weaknesses through the use of different techniques and tools. Attacks are performed by attackers for different reasons like for gaining certain rewards or simply for personal satisfaction. The most commonly performed cyber-attacks on organizations today include Phishing, Password Spraying, Malware, Ransomware, Man-in-the-Middle (MitM), Denial of Service (DoS), and Cross-site Scripting (XSS).

## B. Measures and Security Precautions against Cyber Attacks

Individuals and businesses could benefit from the use of cybersecurity measures as it could provide basic protection to both individuals and businesses from the most common cybersecurity risks. The tools and processes of cybersecurity measures are rather easy to follow. They include usage of strong passwords, a firewall, control of access, security software, intrusion monitoring, and increased awareness.

Most global data breaches (86%) are financially driven [78]. It is thus highly crucial for both individuals and businesses to be proactive in preserving their cybersecurity, considering that a cyber-attack can cost the company millions of dollars [78].

In order that their business data, cash flow and customers are safe online, businesses should employ various types of cybersecurity measures, so that risks from various sources could be averted. Among these sources include internet attacks, user faults, system or software defects and weaknesses, and subvert system or software features [79].

The interview participants were asked concerning their applied measures and security precautions in countering cyber-attacks. According to one of the participants:

*“My organization uses two types of security measures and precautions – the traditional ones and the modern ones. The traditional ones include the common methods like IPS and IDS antivirus. They are used in different platforms. For modern measures, my organization uses anti-malware, DLP solutions and sandbox solutions.”*

On the same question, another interview participant responded by stating the following:

*“In our organization, the main focus is to standardize and adapt all the used measures. The human factor is important as well. We believe that in modern cybersecurity measures, human capital is the most vital element. That is why our organization regularly trains the employees on how to avoid recurrent cyber-attacks.”*

The third participant offered his viewpoint as well. He said:

*“I think that training programs are useful when you want to increase the awareness of personnel towards cybersecurity risks and cyber-attacks. Also, my organization is using security software and programs from different companies. This is to reduce the risks of cyber-attacks. For example, the antivirus and firewall software that my organization is using, was created by different companies. We feel that the risks and threats of cyber-attacks can be reduced this way.”*

For individuals, among the tips that they can follow in preserving their cybersecurity include creating a unique password for each account that they have and updating the password every three months, aside from constantly updating their software to prevent software flaws and to keep their software up to date. According to one participant:

*“It is important that social media users make their account private. Also, they should not reveal their sensitive information in their posts. In fact, I think that social media users need to have knowledge about how to properly use the internet.”*

For protecting business data, among the tips for organizations to follow include the use of secure hardware, data backup and encryption, cybersecurity insurance, security-focused culture, and strong cybersecurity software. One participant accordingly reacted to these tips by saying: *“These steps will help in decreasing the risk and the business can operate in a smooth manner.”*

Prevention is key to reducing the risk of a data breach. Through the use of cybersecurity software, and through having the knowledge of the common attack methods, both individuals and organizations could preserve the privacy of their data and inhibit hackers.

### C. The Future of Cyber Attacks and Related Security Measures

In today’s technology, cybersecurity is an integral element, amounting to approximately \$250 billion in market value [80]. Like the general tech industry, there is a market for cybersecurity in nearly all industries, especially now that the amount of company information being stored online is increasing, and so is the demand for cybersecurity solutions to assure its security.

Leaks and hacks on a large scale have occurred in the last several years, and they have cost companies their customer’s information and also their reputation [76]. For consumers, such occurrences have impaired their sense of safety and their security, and it could even disrupt their lives. Hackers seek leaked information as such information allows them to steal money and sensitive information from people.

In order to get access to lists of user accounts, hackers will go to major websites. However, hackers may face hurdles in their attempts to steal personal information of users from those sites. As stated by one participant of the interview:

*“Our prediction is that cyber-criminals will learn new and innovative ways to attack people, their homes and devices, in their efforts of finding a path to your trustworthy corporate network.”*

Then, another participant added to the discussion by stating: *“Two-factor authorization is now commonly practiced by consumers. There are also new back-end security protocols and security methods. AI is one good example, and I am confident that AI has the ability to disable hacking attempts. Still, we should not say that everything is secure because we all know that hackers will keep finding ways around security structures. Hence, cybersecurity professionals must always be one step ahead.”*

In discussing the topic of Ransomware, a participant said: *“Criminals have been making good profit through Ransomware these past several years. They would block users from using their computers and networks using some malicious software. These users are forced to pay large amounts of money or ransom to these criminals to regain their computer or network use. Unfortunately, we can expect that ransomware attacks will increase.”*

It has been reported that the damages caused by ransomware had amounted to USD \$11.5 billion in 2019 and the amount was expected to increase to USD \$20 billion by 2022 [81].

On phishing attacks, a participant reacted by saying: *“Phishing attacks are increasing and it is increasingly more difficult to detect them. A good example is the MailChimp phishing campaign case. Hackers made use of affected accounts in MailChimp to distribute malicious emails with malware, and considering that MailChimp is a trusted and well-established email marketing provider, the emails were likely to bypass the spam filters and enter the inbox of the unwary receiver.”*

Some experts agreed with the prediction that cybercrimes would cause financial damages amounting to \$6 trillion by the end of 2022 [82]–[84]. Also, a cyber-attack would be expected to occur in every 11 seconds, as compared to one attack in every 19 seconds reported in 2019. Comparatively, in 2016, one cyber-attack occurred every 40 seconds [82]–[84].

The COVID 19 outbreak has forced many to work remotely. On this matter, a participant stated the following:

*“COVID 19 has sped up the digital transformation of organizations and many employees have shifted to working from home. The problem with this situation is that, when working from home or remotely, employees are not safeguarded by corporate firewalls. Hackers could exploit the vulnerabilities that they discover in the gaps between people, their devices, and the corporate network. We can see that many employees are required to establish workloads in Cloud, and so, cloud-based security techniques are crucial in curbing the failing cybersecurity landscape. It is now necessary to work with cloud-native Identity and Access Management (IAM).”*

As stated by another participant: *“For those who maintain cloud-based security, they have to have adequate ability in managing the infrastructure with the use of structured programs. We know that networks and application tiers are short-lived, and so, for any organization, their most crucial asset is probably their own data and the data of their customers. Therefore, I think that data-security on the cloud will be a main theme in the future.”*

One participant viewed: *“cybersecurity as a highly viable career path.”* In fact, it was predicted that the number of vacant cybersecurity jobs would increase by 35%, considering the increase in cyber-attacks and the shifts in tactics used by cybercriminals [85, 86]. In other words, cybersecurity is likely to become a sound career choice in the long run.

In terms of AI application, all interviewees agreed that it can effectively improve the security of cyberspace. Also, the options of AI systems should be fully used so that the most optimal level of cybersecurity could be achieved. One participant indicated that: *“the use of AI can allow the discovery of new and refined transformations in attack flexibility.”* Notably, the old technology concentrates on the past and is mostly focusing on identified attackers and attacks. This can lead to the formation of blind spots in the detection of uncommon behavior in new attacks. For instance, privileged activity in an intranet can be monitored, and any discernible alteration in privileged access operations can signify a likely internal threat. For successful detection, the machine will strengthen the validity of the actions and increase its sensitivity for the detection of equivalent patterns in the future.

According to one participant: *“AI facilitates machine*

learning. It also helps the machine in detecting irregularities more effectively, and also in increasing the accurateness of operations. AI is particularly useful in dealing with more sophisticated cyber-attacks because the approaches used by hackers are increasingly more innovative these days.”

One participant indicated that: “the use of AI can increase network security, particularly through the ability of AI in detecting attacks and in responding to breaches.” For security groups, they could be overwhelmed by the amount of security alerts that they receive on a daily basis. Hence, the automatic detection and response towards threats can decrease their workload significantly. Also, they could more effectively detect the threats with the use of AI.

The formation and transmission of colossal amounts of security data on a daily basis would gradually impair the ability of security experts in quickly and reliably tracking and identifying the attack factors [87]. With the use of AI, the monitoring and detection of doubtful behavior could be expanded. This allows the network security personnel to effectively and promptly react to new situations.

The majority of participants agreed with the fact that in improving its response to threats, AI security systems have the ability to learn over time. Utilizing AI will facilitate the detection of threats following the application behavior and the entire network activity. The AI security system studies the standard network traffic and behavior, and over time, AI creates a reference point on what constitutes a normal pattern, and any divergences from the norm can be identified to determine attacks.

Considering the viewpoints of the participants in the interviews, it is clear that cybersecurity experts have to be consistently ahead of the cybercriminals. Also, the use of AI techniques can effectively improve cyberspace security. The options of AI systems should be fully utilized in order to achieve the most optimal level of cybersecurity. Lastly, remote working requirements owing to the COVID 19 outbreak have resulted in the value of cloud-based security techniques in protecting and improving the cybersecurity landscape.

## V. CONCLUSION AND FUTURE WORK

This paper reviews the topic of cyber-attacks focusing on the types and weapons of cyber-attacks, measures and security precautions against cyber-attacks, and the projections of experts on the future of cyber-attacks and the associated security measures.

Interviews were carried with several study participants, delving into the subjects of cybersecurity and cyber-attacks. The information obtained from the participants shows dramatic development of information technology within the past several years. It was found that albeit the presence of precautionary tools, cyber attackers are still successful in breaking the fire-wall systems, resulting in physical and non-physical damages. Victims of cyber-attacks, especially firms, could lose their reputation as well. Notably, cyber-attack risks and threats appear to increase in tandem with information technology development.

Taking into consideration the present situation of data breaches, ransomware attacks, in addition to the concerns

towards the impact of new technologies like AI, and the constantly evolving threats; it is a critical duty of cybersecurity experts to consistently provide the most updated best practices and tools of cybersecurity so that users could consistently avert cyber threats. Equally, employee awareness and expertise on cybersecurity issues need to be encouraged and cultivated via continuous training programs.

Further, scientific research should focus more on security adjustments and measures required by different business sectors and corporations towards remote work that tends to become a necessity today; especially with the spread of Corona pandemic. More focus on individual safety characteristics such as awareness, attitude, behavior and compliance is called for, and research is required to quantify these primary quality indicators. Based on past findings, the human factor appears the key to the progress of information security, but has not been adequately explored.

## REFERENCES

- [1] S. Tan, P. Xie, J. M. Guerrero, J. C. Vasquez, Y. Li, and X. Guo, “Attack detection design for dc microgrid using eigenvalue assignment approach,” *Energy Reports*, vol. 7, pp. 469–476, 2021.
- [2] M. A. Judge, A. Manzoor, C. Maple, J. J. Rodrigues, and S. ul Islam, “Price-based demand response for household load management with interval uncertainty,” *Energy Reports*, vol. 7, pp. 8493–8504, 2021.
- [3] Y. Li and Q. Liu, “A comprehensive review study of cyber-attacks and cyber security; emerging trends and recent developments,” *Energy Reports*, vol. 7, pp. 8176–8186, 2021.
- [4] G. Aghajani and N. Ghadimi, “Multi-objective energy management in a micro-grid,” *Energy Reports*, vol. 4, pp. 218–225, 2018.
- [5] I. Priyadarshini, R. Kumar, R. Sharma, P. K. Singh, and S. C. Satapathy, “Identifying cyber insecurities in trustworthy space and energy sector for smart grids,” *Computers & Electrical Engineering*, vol. 93, p. 107204, 2021.
- [6] J. Al-Gasawneh, A. AL-Hawamleh, A. Alorfi, and G. Al-Rawashde, “Moderating the role of the perceived security and endorsement on the relationship between perceived risk and intention to use the artificial intelligence in financial services,” *International Journal of Data and Network Science*, vol. 6, no. 3, pp. 743–752, 2022.
- [7] H. Akhavan-Hejazi and H. Mohsenian-Rad, “Power systems big data analytics: An assessment of paradigm shift barriers and prospects,” *Energy Reports*, vol. 4, pp. 91–100, 2018.
- [8] N. Li, C. Tsigkanos, Z. Jin, Z. Hu, and C. Ghezzi, “Early validation of cyber-physical space systems via multi-concerns integration,” *Journal of Systems and Software*, vol. 170, p. 110742, 2020.
- [9] J. Shin, J.-G. Choi, J.-W. Lee, C.-K. Lee, J.-G. Song, and J.-Y. Son, “Application of stpa-safesec for a cyber-attack impact analysis of npps with a condensate water system test-bed,” *Nuclear Engineering and Technology*, vol. 53, no. 10, pp. 3319–3326, 2021.
- [10] M. Snehi and A. Bhandari, “Vulnerability retrospection of security solutions for software-defined cyber-physical system against ddos and iot-ddos attacks,” *Computer Science Review*, vol. 40, p. 100371, 2021.
- [11] A. A. Jamal, A.-A. M. Majid, A. Konev, T. Kosachenko, and A. Shelupanov, “A review on security analysis of cyber physical systems using machine learning,” *Materials Today: Proceedings*, 2021.
- [12] A. M. Hawamleh and A. Ngah, “An adoption model of mobile knowledge sharing based on the theory of planned behavior,” *Journal of Telecommunication, Electronic and Computer Engineering (JTEC)*, vol. 9, no. 3-5, pp. 37–43, 2017.
- [13] S. W. Brenner, *Cyberthreats: The emerging fault lines of the nation state*. Oxford University Press, 2009.
- [14] B. Alhayani, S. T. Abbas, D. Z. Khutar, and H. J. Mohammed, “Best ways computation intelligent of face cyber attacks,” *Materials Today: Proceedings*, 2021.

- [15] A. Hawamleh, A. S. M. Alorfi, J. A. Al-Gasawneh, and G. Al-Rawashdeh, "Cyber security and ethical hacking: The importance of protecting user data," *Solid State Technology*, vol. 63, no. 5, pp. 7894–7899, 2020.
- [16] S. Cheung, U. Lindqvist, and M. W. Fong, "Modeling multistep cyber attacks for scenario recognition," in *Proceedings DARPA Information Survivability Conference And Exposition*, vol. 1. IEEE, 2003, pp. 284–292.
- [17] I. Frank and E. Odunayo, "Approach to cyber security issues in nigeria: challenges and solution," *International Journal of Cognitive Research in science, engineering and education*, vol. 1, no. 1, pp. 100–110, 2013.
- [18] P. Seemma, S. Nandhini, and M. Sowmiya, "Overview of cyber security," *International Journal of Advanced Research in Computer and Communication Engineering*, vol. 7, no. 11, pp. 125–128, 2018.
- [19] C. O. K. CLN, E. I. C.-K. CLN, I. A. A. O. CLN, and B. A. U. CLN, "Issues on information systems, icts, cyber-crimes, cyber security, cyber ethics, and national security in nigeria: Librarians' research," *Library Philosophy and Practice*, pp. 1–19, 2020.
- [20] S. Al-Emadi, A. Al-Mohannadi, and F. Al-Senaid, "Using deep learning techniques for network intrusion detection," in *2020 IEEE International Conference on Informatics, IoT, and Enabling Technologies (ICIoT)*. IEEE, 2020, pp. 171–176.
- [21] L. Griffin, "The effectiveness of cybersecurity awareness training in reducing employee negligence within department of defense (dod) affiliated organizations-qualitative exploratory case study," Ph.D. dissertation, Capella University, 2021.
- [22] T. Bhardwaj, H. Upadhyay, and L. Lagos, "Deep learning-based cyber security solutions for smart-city: Application and review," *Artificial Intelligence in Industrial Applications*, pp. 175–192, 2022.
- [23] B. Cashell, W. D. Jackson, M. Jickling, and B. Webel, "The economic impact of cyber-attacks," *Congressional research service documents, CRS RL32331 (Washington DC)*, vol. 2, 2004.
- [24] F. Skopik, G. Settanni, and R. Fiedler, "A problem shared is a problem halved: A survey on the dimensions of collective cyber defense through security information sharing," *Computers & Security*, vol. 60, pp. 154–176, 2016.
- [25] K. Thakur, M. L. Ali, S. Kopecky, A. Kamruzzaman, and L. Tao, "Connectivity, traffic flow and applied statistics in cyber security," in *2016 IEEE International Conference on Smart Cloud (SmartCloud)*. IEEE, 2016, pp. 295–300.
- [26] S. Demirkan, I. Demirkan, and A. McKee, "Blockchain technology in the future of business cyber security and accounting," *Journal of Management Analytics*, vol. 7, no. 2, pp. 189–208, 2020.
- [27] W. Steingartner, D. Galinec, and A. Kozina, "Threat defense: Cyber deception approach and education for resilience in hybrid threats model," *Symmetry*, vol. 13, no. 4, p. 597, 2021.
- [28] O. T. Soyoye and K. C. Stefferud, "Cybersecurity risk assessment for california's smart inverter functions," in *2019 IEEE CyberPELS (CyberPELS)*. IEEE, 2019, pp. 1–5.
- [29] M. Lezzi, M. Lazoi, and A. Corallo, "Cybersecurity for industry 4.0 in the current literature: A reference framework," *Computers in Industry*, vol. 103, pp. 97–110, 2018.
- [30] W. A. Al-Khater, S. Al-Maadeed, A. A. Ahmed, A. S. Sadiq, and M. K. Khan, "Comprehensive review of cybercrime detection techniques," *IEEE Access*, vol. 8, pp. 137 293–137 311, 2020.
- [31] N. Setiawan, V. C. E. Tarigan, P. B. Sari, Y. Rossanty, M. Nasution, and I. Siregar, "Impact of cybercrime in e-business and trust," *Int. J. Civ. Eng. Technol*, vol. 9, no. 7, pp. 652–656, 2018.
- [32] T. Holt and A. Bossler, *Cybercrime in progress: Theory and prevention of technology-enabled offenses*. Routledge, 2015.
- [33] R. Anderson, C. Barton, R. Böhme, R. Clayton, M. J. Van Eeten, M. Levi, T. Moore, and S. Savage, "Measuring the cost of cybercrime," in *The economics of information security and privacy*. Springer, 2013, pp. 265–300.
- [34] S. Gordon and R. Ford, "On the definition and classification of cybercrime," *Journal in computer virology*, vol. 2, no. 1, pp. 13–20, 2006.
- [35] A. C. Moise *et al.*, "A few comments on the council of europe convention on cybercrime," *Jurnalul de Drept si Stiinte Administrative*, vol. 2, no. 8, pp. 28–38, 2017.
- [36] N. C. Hampson, "Hacktivism: A new breed of protest in a networked world," *BC Int'l & Comp. L. Rev.*, vol. 35, p. 511, 2012.
- [37] T. U. Rehman, "Psychosocial aspects of cybercrime victimization in pakistan," in *Handbook of Research on Applied Social Psychology in Multiculturalism*. IGI Global, 2021, pp. 192–211.
- [38] D. Shivpuri, "Cyber crime: Are the law outdated for this type of crime," *International Journal of Research in Engineering, Science and Management*, vol. 4, no. 7, pp. 44–49, 2021.
- [39] A. Sarmah, R. Sarmah, and A. J. Baruah, "A brief study on cyber crime and cyber law's of india," *International Research Journal of Engineering and Technology (IRJET)*, vol. 4, no. 6, pp. 1633–1640, 2017.
- [40] M. Abomhara and G. M. Kjøen, "Cyber security and the internet of things: vulnerabilities, threats, intruders and attacks," *Journal of Cyber Security and Mobility*, pp. 65–88, 2015.
- [41] C. Ventures, "2019 official annual cybercrime report," in *Recuperado el*. Herjavec Group, 2019.
- [42] R. Fisher, C. Porod, and S. Peterson, "Motivating employees and organizations to adopt a cybersecurity-focused culture," *Journal of Organizational Psychology*, vol. 21, no. 1, pp. 114–131, 2021.
- [43] A. Al-Marghilani, "Comprehensive analysis of iot malware evasion techniques," *Engineering, Technology & Applied Science Research*, vol. 11, no. 4, pp. 7495–7500, 2021.
- [44] A. Goel, D. K. Sharma, and K. D. Gupta, "Leobat: Lightweight encryption and otp based authentication technique for securing iot networks," *Expert Systems*, vol. 39, no. 5, p. e12788, 2022.
- [45] Y. E. Suzuki and S. A. S. Monroy, "Prevention and mitigation measures against phishing emails: a sequential schema model," *Security Journal*, vol. 35, no. 4, pp. 1162–1182, 2022.
- [46] B. B. Gupta, A. Tewari, A. K. Jain, and D. P. Agrawal, "Fighting against phishing attacks: state of the art and future challenges," *Neural Computing and Applications*, vol. 28, no. 12, pp. 3629–3654, 2017.
- [47] M. Conti, N. Dragoni, and V. Lesyk, "A survey of man in the middle attacks," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 3, pp. 2027–2051, 2016.
- [48] K. M. Prasad, A. R. M. Reddy, and K. V. Rao, "Dos and ddos attacks: defense, detection and traceback mechanisms-a survey," *Global Journal of Computer Science and Technology*, 2014.
- [49] S. Shalini and S. Usha, "Prevention of cross-site scripting attacks (xss) on web applications in the client side," *International Journal of Computer Science Issues (IJCSI)*, vol. 8, no. 4, p. 650, 2011.
- [50] M. Souppaya, K. Scarfone *et al.*, "Guide to malware incident prevention and handling for desktops and laptops," *NIST Special Publication*, vol. 800, p. 83, 2013.
- [51] A. Sheikh, "Trojans, backdoors, viruses, and worms," in *Certified Ethical Hacker (CEH) Preparation Guide*. Springer, 2021, pp. 49–69.
- [52] S. Sharma, "Design and implementation of malware detection scheme," *International Journal of Computer Network & Information Security*, vol. 10, no. 8, 2018.
- [53] M. Rai and H. Mandoria, "A study on cyber crimes cyber criminals and major security breaches," *Int. Res. J. Eng. Technol.*, vol. 6, no. 7, pp. 1–8, 2019.
- [54] B. Narwal, A. K. Mohapatra, and K. A. Usmani, "Towards a taxonomy of cyber threats against target applications," *Journal of Statistics and Management Systems*, vol. 22, no. 2, pp. 301–325, 2019.
- [55] I. A. Chesti, M. Humayun, N. U. Sama, and N. Jhanjhi, "Evolution, mitigation, and prevention of ransomware," in *2020 2nd International Conference on Computer and Information Sciences (ICIS)*. IEEE, 2020, pp. 1–6.
- [56] K. K. Gagneja, "Knowing the ransomware and building defense against it-specific to healthcare institutes," in *2017 Third International Conference on Mobile and Secure Services (MobiSecServ)*. IEEE, 2017, pp. 1–5.
- [57] M. Papoutsakis, K. Fysarakis, G. Spanoudakis, S. Ioannidis, and K. Koloutsou, "Towards a collection of security and privacy patterns," *Applied Sciences*, vol. 11, no. 4, p. 1396, 2021.
- [58] S. Boonkrong, "Methods and threats of authentication," in *Authentication and Access Control*. Springer, 2021, pp. 45–70.

- [59] A. Kanta, S. Coray, I. Coisel, and M. Scanlon, "How viable is password cracking in digital forensic investigation? analyzing the guessability of over 3.9 billion real-world accounts," *Forensic Science International: Digital Investigation*, vol. 37, p. 301186, 2021.
- [60] R. Beno and R. Poet, "Hacking passwords that satisfy common password policies: Hacking passwords," in *13th International Conference on Security of Information and Networks*, 2020, pp. 1–3.
- [61] V. Nithya, S. L. Pandian, and C. Malarvizhi, "A survey on detection and prevention of cross-site scripting attack," *International Journal of Security and Its Applications*, vol. 9, no. 3, pp. 139–152, 2015.
- [62] A. M. K. Alhawamleh, "Web based english placement test system (elpts)," Ph.D. dissertation, Universiti Utara Malaysia, 2012.
- [63] A. Raman, S. Kaushik *et al.*, "A comprehensive study of contemporary tools and techniques in the realm of cyber security," *IITM Journal of Management and IT*, vol. 7, no. 1, pp. 108–120, 2016.
- [64] J. L. Duffany, "Computer security," in *Computer and Network Security Essentials*. Springer, 2018, pp. 3–20.
- [65] K. Kallepalli and U. B. Chaudhry, "Intelligent security: Applying artificial intelligence to detect advanced cyber attacks," in *Challenges in the IoT and Smart Environments*. Springer, 2021, pp. 287–320.
- [66] M. Chakraborty and M. Singh, "Introduction to network security technologies," in *The "Essence" of Network Security: An End-to-End Panorama*. Springer, 2021, pp. 3–28.
- [67] H. Tabrizchi and M. Kuchaki Rafsanjani, "A survey on security challenges in cloud computing: issues, threats, and solutions," *The journal of supercomputing*, vol. 76, no. 12, pp. 9493–9532, 2020.
- [68] M. Becher, F. C. Freiling, J. Hoffmann, T. Holz, S. Uellenbeck, and C. Wolf, "Mobile security catching up? revealing the nuts and bolts of the security of mobile devices," in *2011 IEEE Symposium on Security and Privacy*. IEEE, 2011, pp. 96–111.
- [69] R. P. Jover, "Security analysis of sms as a second factor of authentication," *Communications of the ACM*, vol. 63, no. 12, pp. 46–52, 2020.
- [70] M. F. Mushtaq, S. Jamel, A. H. Disina, Z. A. Pindar, N. S. A. Shakir, and M. M. Deris, "A survey on the cryptographic encryption algorithms," *International Journal of Advanced Computer Science and Applications*, vol. 8, no. 11, 2017.
- [71] M. B. Yassein, S. Aljawarneh, E. Qawasmeh, W. Mardini, and Y. Khamayseh, "Comprehensive study of symmetric key and asymmetric key encryption algorithms," in *2017 international conference on engineering and technology (ICET)*. IEEE, 2017, pp. 1–7.
- [72] N. G. Kumar and K. K. Rao, "Hash based approach for providing privacy and integrity in cloud data storage using digital signatures," *International Journal of Computer Science and Information Technologies*, vol. 5, no. 6, pp. 8074–8078, 2014.
- [73] D. Hofheinz and T. Jager, "Tightly secure signatures and public-key encryption," *Designs, Codes and Cryptography*, vol. 80, no. 1, pp. 29–61, 2016.
- [74] J. Malzac, "Leveraging domestic law against cyberattacks," *Nat'l Sec. L. Brief*, vol. 11, p. 1, 2021.
- [75] B. Pranggono and A. Arabo, "Covid-19 pandemic cybersecurity issues," *Internet Technology Letters*, vol. 4, no. 2, p. e247, 2021.
- [76] F. Schlackl, N. Link, and H. Hoehle, "Antecedents and consequences of data breaches: A systematic review," *Information & Management*, p. 103638, 2022.
- [77] P. Langlois, "2020 data breach investigations report," *Verizon*, 2020.
- [78] M. Jartelius, "The 2020 data breach investigations report—a cso's perspective," *Network Security*, vol. 2020, no. 7, pp. 9–12, 2020.
- [79] K. Raghavan, M. S. Desai, and P. Rajkumar, "Managing cybersecurity and ecommerce risks in small businesses," *Journal of management science and business intelligence*, vol. 2, no. 1, pp. 9–15, 2017.
- [80] P. Lorenzo, F. Stefano, A. Ferreira, and P. Carolina, "Artificial intelligence and cybersecurity: Technology, governance and policy challenges," 2021.
- [81] H. Lee and K.-S. Choi, "Interrelationship between bitcoin, ransomware, and terrorist activities: Criminal opportunity assessment via cyber-routine activities theoretical framework," *Victims & Offenders*, vol. 16, no. 3, pp. 363–384, 2021.
- [82] Y. Perwej, S. Q. Abbas, J. P. Dixit, N. Akhtar, and A. K. Jaiswal, "A systematic literature review on the cyber security," *International Journal of scientific research and management*, vol. 9, no. 12, pp. 669–710, 2021.
- [83] S. Gangwar and V. Narang, "A survey on emerging cyber crimes and their impact worldwide," in *Research Anthology on Combating Cyber-Aggression and Online Negativity*. IGI Global, 2022, pp. 1583–1595.
- [84] N. Hassan, *Ransomware Revealed*. Springer, 2019.
- [85] W. Crumpler and J. A. Lewis, *The cybersecurity workforce gap*. Center for Strategic and International Studies (CSIS) Washington, DC, USA, 2019.
- [86] A. Kanaan, A. AL-Hawamleh, A. Abulfaraj, H. Al-Kaseasbeh, and A. Alorfi, "The effect of quality, security and privacy factors on trust and intention to use e-government services," *International Journal of Data and Network Science*, vol. 7, no. 1, pp. 185–198, 2023.
- [87] T. C. Truong, I. Zelinka, J. Plucar, M. Čandík, and V. Šulc, "Artificial intelligence and cybersecurity: Past, presence, and future," in *Artificial intelligence and evolutionary computations in engineering systems*. Springer, 2020, pp. 351–363.

# BERT Model-based Natural Language to NoSQL Query Conversion using Deep Learning Approach

Kazi Mojammel Hossen<sup>1</sup>, Mohammed Nasir Uddin<sup>2</sup>, Minhazul Arefin<sup>3</sup>, Md Ashraf Uddin<sup>4</sup>  
Department of CSE, Jagannath University  
Dhaka, Bangladesh<sup>1,2,3,4</sup>

**Abstract**—Databases are commonly used to store complex and distinct information. With the advancement of the database system, non-relational databases have been used to store a vast amount of data as traditional databases are not sufficient for making queries on a wide range of massive data. However, storing data in a database is always challenging for non-expert users. We propose a conversion technique that enables non-expert users to access and filter data as close to human language as possible from the NoSQL database. Researchers have already explored a variety of technologies in order to develop more precise conversion procedures. This paper proposed a generic NoSQL query conversion learning method to generate a Non-Structured Query Language from natural language. The proposed system includes natural language processing-based text preprocessing and the Levenshtein distance algorithm to extract the collection and attributes if there were any spelling errors. The analysis of the result shows that our suggested approach is more efficient and accurate than other state-of-the-art methods in terms of bilingual understudy scoring with the WikiSQL dataset. Additionally, the proposed method outperforms the existing approaches because our method utilizes a bidirectional encoder representation from a transformer multi-text classifier. The classifier process extracts database operations that might increase the accuracy. The model achieves state-of-the-art performance on WikiSQL, obtaining 88.76% average accuracy.

**Keywords**—Natural language processing; NoSQL query; BERT model; Levenshtein distance algorithm; artificial neural network

## I. INTRODUCTION

In today's digital age, non-relational databases are utilized in almost every industry to store information. Non-Structured Query Language (NoSQL) databases [1], [2] are increasingly being used for large-scale data sets, search engines, and real-time web applications [3]. Nowadays, NoSQL databases work as an alternative to relational databases [4] and other conventional databases [5].

With the growth of technology, NoSQL databases stores a large amount of data in document stores, key-value data stores, wide-column stores, and Graph stores. As opposed to relational databases, MongoDB, CouchDB, Cassandra, etc are designed on the architecture of distributed systems to store massive data [6]. Many organizations are gradually looking into approaches to understand and analyze this enormous unstructured data. The current approaches to data management, organization, and storage are being changed by "Big Data" [7]. In particular, "Big Data," an open source framework used to store vast amounts of structured, unstructured, and semi-structured data [8]. So, Normal users require knowledge of the query syntax and table schema to access and store a large amount of data. However, finding a reliable approach to generate the NoSQL

query from Natural Language (English) is challenging. Using NoSQL approach, amateur users can interact with the database system. The model facilitates communication between humans and computers without recalling the query syntax method for the non-relational databases. Natural Language Processing (NLP) [9], [10], [11] is a branch of linguistics, information engineering, computer science, and artificial intelligence that studies how computers and humans interact with Natural Language [12]. Traditional machine translation is applied to translate the text from one language to another by NLP [13].

This research aims to develop a feasible tool for searching databases where natural language can be used without needing complex database queries that are developed by expertise. Generating NoSQL from natural language has wide range of applications. Tools with AI knowledge [14] such as Google Assistant or Alexa use the NLIDB system for non-technical users. Filling out a lengthy online form can be tedious and users might need to navigate through the screen, scroll, look up values in the scroll box, and so on whereas with NLIDB, the users need to type a question similar to a sentence. Consequently, such a tool has a wide range of usage and applications. NoSQL approach has been researched both in academia as well as in industry [15]. In this paper, we implement a Neural Machine translation model which consists of four steps. First, we have used a Natural Language Tool-Kit for performing text preprocessing. Secondly, attributes are collected and extracted using Levenshtein Distance (LD) [16], [17] algorithm. Thirdly, we have used a bidirectional encoder representations from BERT Transformers Model-based multi-text classification [18] to extract the operations including find, insert, update and remove. The last step of the proposed approach is generating query.

Many research works have used WIKISQL dataset for conversation Natural Language to Structured Query Language. The BERT Model generates the NoSQL operational command from the WIKISQL task. The contribution of this research paper can be summarized as follows:

- Designing several algorithms to come up with a standard machine translation model for converting Natural Language into NoSQL queries.
- To resolve the syntax errors for primitive users using Levenshtein Distance algorithm that can extract the collection and attributes from the text even if any users make spelling mistakes or utilize synonyms.
- To employ the latest contextual word representation BERT transformer model to extract the operations with a higher accuracy rate.

The remainder of the paper is organized as follows: Related works conducting with the same and different technologies by other researchers are illustrates in Section II. Section III describes the proposed methodology and work flow. Section IV shows experiment evaluation and result of the proposed system. Conclusions with the future expansion are detailed in Section V.

## II. RELATED WORK

Research in Natural Language for non-relational databases has started as far back as the twenty century. Since the interest in Natural Language Processing has continued tremendously. In the early 1970's LUNAR [19], the first Natural Language Interface for the relational database (NLIDB) has introduced to the researcher. LUNAR was a Question Answering (QA) system connected with the moon rock sample database. The information of rock samples brought back from the moon was used to make the LUNAR database. NLP to NoSQL query conversion field has very little research on it. This section discusses various works on Natural Language to query conversion.

In 2021, Minhazul et al. [20] suggested a machine learning-based NLP2SQL translation system. They used the Naive Bayes algorithm for command extraction and decision tree regression for condition extraction. Their proposed method lack accuracy because of using the bag of words technique in the derivation of condition from SQL. An advance deep learning solution can mitigate this problem. On the other hand, they can use the neural translation technique for this machine translation approach. The system can use the statistical translation method also.

Mallikarjun et al. [21] proposed an automated NLP-based text processing approach. Their approach can successfully convert an excel datasheet into a DBMS. Their system has a user authentication system that prevents unwanted users. The system has a limitation of 16,384 columns and 1,048,576 rows for an excel worksheet. This data may be massive for average purposes but not enough for big data.

An Intelligent processing system in a document-based NoSQL database had proposed by Benymol et al. [22] in 2021. They used state-of-the-art algorithms and technologies to convert text into NoSQL. They used different types of TF-IDF schemes for information retrieval, machine learning algorithm for modeling, and hyper parameter tuning for model selection. The system may have vulnerability in stream and batch data on the Big Data processing platform. The proposed model also has a problem with dynamic processing strategies. In this stage, the system fails to find any possible solution.

Fatma et al. [23] proposed an automatic UML/OCL model for the NoSQL database converter. Their system mainly focuses on the big data platform. Because there is wide use of NoSQL database in the big data platform. After creating the NoSQL database, the system automatically checks the OCL constraints of the model. There are different types of NoSQL databases and a maximum of them have a problem with integrity constraint checking. For this, it is the most challenging task in the system.

In [24] M. T. Majeed et al. have designed a fully automated framework that, using an AI technique, can recognize

keywords, symbols, attributes, values, and relationships among various types of queries. Additionally, they proposed a graphical user interface where users could enter NL queries and have a NoSQL query created from those queries. For complex queries, the proposed framework didn't offer a solution.

S. Mondal et al. [25] introduced a query-response model that can respond to a variety of queries, including assertive, interrogative, imperative, compound, and complicated forms. This NoSQL system's primary task is to retrieve knowledge data from the default MongoDB database. This paper didn't give any solution of time-related query such as "What is the age of x after 10 years".

T. Pradeep et al. [26] presented a Deep Learning based approach that converts English questions to MongoDB queries. They applied an encoder-Decoder machine-translation method for this conversion. The encoder turns the NLQ text input into a vector and sends it to the decoder. The decoder uses a deep neural network to predict NoSQL queries. Their system uses ten different deep learning models to handle ten types of MongoDB queries. One solution is the best possible answer for this problem.

Sebastian Blank et al. [27] suggested an end-to-end Question Answering (QA) system. It allows a user to ask a question in natural language on the Elasticsearch database. They solve the homogeneous operation problem of the database by using policy-based reinforcement learning. For that, they used Facebook's bAbI Movie Dialog dataset. They also design a KBQueryBot, an agent of translating a natural language query into the domain-specific query language based on a sequence-to-sequence model [28]. It gives every single answer with the help of an external knowledge base.

Some classic NLIDB systems can solve the spelling corrections of misspelled words automatically [29]. The module gives the interface between computer and user by the database query language. Consequently, they discuss the overall system architecture of the NLIDB, some implementation details, and experimental results. The proposed work only focuses on automatic spelling and grammar correction.

Z. Farooqui et al. [30] recommended the conversion of English to SQL. For example, their system converts English questions or text queries into SQL queries. Later it will be operated on databases. Their suggested technique and method are generic and smooth. It can handle both small and large applications for generic NLIDB systems. There are four types of input NLQ text Normal, Linear Disjoint, Linear Coincident, and Non-Linear Model. It focuses on simple SQL query clauses such as SELECT, FROM, WHERE, and JOIN. Their system can handle complex queries resulting from ambiguous NL queries.

Tanzim Mahmud et al. [31] proposed a system based on Context-Free-Grammar (CFG). Any input token of appropriate terminals found in the input NLQ will replace the corresponding attribute in the relational table or applicable operators of SQL. The interface can configure easily and automatically by the user. It relies on the Metadata set and Semantic sets for tables and attributes. It can handle ambiguities in the input NLQ. For example, the system can solve the same attribute name clashing problem within a table. The limitation of the

proposed CFG system can only convert limited queries. Other than that, the system is dependent on a specific language.

Xiaojun Xu et al. proposed SQLNet [32], an NLP to SQL conversion approach which is order-independent and alternative to the traditional sketch based program synthesis approaches [33]. Failing an order input NLQ text is not a problem in that case. It uses a sequence-to-set model, which is a column attention mechanism that generates SQL queries. It represents pseudo-tasks with the help of a function of relevance and works on the WikiSQL task.

Victor Zhong et al. [34] thought about the availability of the query ground truth (intermediate labels) and database response. They proposed Seq2SQL, a modular approach that translates NLQ into SQL queries. Their suggested system also generalizes across different table schemas. There are three modules in Seq2SQL. The first module tries to identify an aggregator function like MIN() or MAX(). The second module extracts column names from NLQ and uses them as a select operator. Both modules worked on question-answer pairs. The third module extracts condition or where-clause from NLQ. There is a possibility to swap between arguments in the WHERE clause. This ambiguity problem could solve by policy-based reinforcement learning [35] in question-answer pairs.

### III. PROPOSED METHODOLOGY

The main concept behind this method is to transform Natural Language (NL) into Non-Structured Query Language (NSQL) using Natural Language Tool Kit (NLTK) and Deep Learning Model. The concept and its description are formalized in the following sections. The proposed architecture is shown in Fig. 1.

#### A. Input Natural Language Query (NLQ)

NLQ consists of only the normal terms of user's language, without any special format or syntax. Natural language query (NLQ) in English is given as input. This input text will be processed for getting information and later converted into NoSQL queries.

1) *Text preprocessing*: Since the inventory of individual words, text can take many forms, ranging from sentences to many paragraphs with special letters. In NLP the text preprocessing is an important task and the first step in the preprocessing to building a model. It is a data mining technique that transforms plain text into a machine-readable format. Real-world data is frequently inadequate, inconsistent, or deficient in specific behaviors and is likely to contain various errors. This step is needed for transferring input text from human language to machine-readable format for further processing.

In this paper, we have used NLTK for text preprocessing. The NLTK is the most widely used and well-known of the NLP libraries in the Python ecosystem. It is used for all sorts of tasks from lowercase conversion to tokenization, removing escape words, part of speech tagging, and beyond. Input text will be processed for getting information from Natural language Query input. From the processed text, the system will extract collection, attribute, and operation for making a NoSQL query.

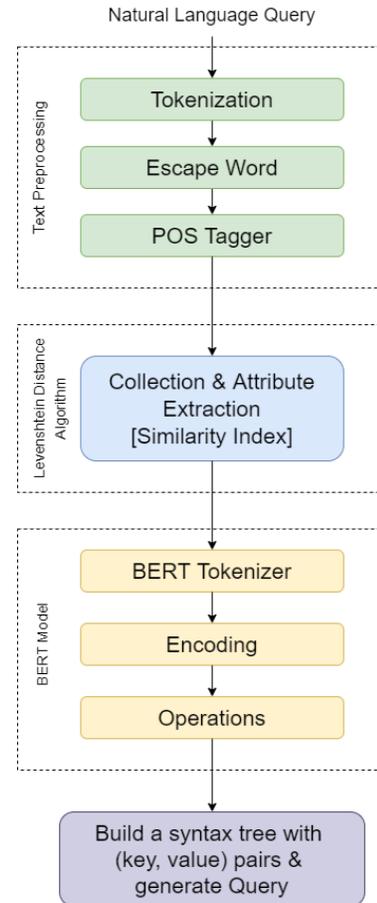


Fig. 1. Proposed methodology

2) *Lowercase conversion*: Lowercase conversion is the first step of text preprocessing. In this step, the input NLQ is converted into a lower case format. Although the uppercase or lowercase forms of words are supposed to have no difference, all the uppercase characters usually changed into lowercase forms before the classification.

3) *Tokenization*: Tokenization splits the natural language query, phrase, string, or entire text document into smaller units such as individual words or tokens. The former Sentence Boundary Disambiguation (SBD) is often used to form a list of individual sentences. It depends on a pre-trained, language-specific algorithm similar to the Punkt Models from NLTK. The text divides into a list of words using an unsupervised algorithm to form a model for abbreviated words. For the English language, a pre-trained Punkt tokenizer includes in the NLTK data package.

4) *Removing escape words*: Escape or extra words are the words that are frequently appeared within the text without having more information or content. So, the escape words are removed because they are not needed in the analysis of the query. For the purpose of building queries, several sets of escape words have been developed. In this paper, we proposed a new set of escape words. Auxiliary verbs and prepositions are mainly used in this context as escape words such as 'a', 'an', 'the', 'is', 'of', 'with', 'to', 'for', 'and', 'all', etc.

TABLE I. ESCAPE WORDS

Escape Words	With Escape Words	Without Escape Words
all, the	Find all the students	Find students
is, the, of, all	What is the name of all student?	What name student
the, and	Insert the student name x and age 20	Insert student name x age 20

Beside Table I, this step eliminates punctuation from the input natural query. The detailed process of eliminating escape words is illustrated in Algorithm 1.

**Algorithm 1: Removing Extra Words**

**Input:** I = Input words and; E = List of Extra Words  
**Output:** L: List of words after removing extra words  
 $c_w = \text{CountWord}(I)$   
**for**  $c_w \in C_W$  **do**  
     $I = I[c]$  **TOKENS** = Tokenization(I)  
    **for**  $l \in \text{TOKENS}$  **do**  
        **TOKEN** = EMPTY  
        **if**  $l \notin E$  **then**  
            |  $\text{PUSH}(\text{TOKEN}, t)$   
        **end**  
    **end**  
     $\text{PUSH}(L, \text{TOKEN})$   
**end**  
**return** L

Removing escape words is a simple but essential aspect of many text mining applications cause it reduces memory overhead. It can reduce noise and false positives. This method can potentially improve the power of prediction in any text mining application.

5) *Parts of Speech (PoS) tagger:* PoS tagging helps in text-to-speech conversion, information retrieval, and word sense disambiguation. It's used for the classification of words in their PoSs and labeling them according to the tagset. The collection of tags used for PoS tagging is tagset. PoS tagging is also referred to as word classes or lexical categories. However, all PoS tags aren't necessary to analyze. All PoS tagging attributes are provided by the NLTK toolkit. The PoSs must be defined as the following:

- Noun Tags = ['NN', 'NNS', 'NNP', 'NNPS']
- Adjective Tags = ['JJ', 'JJS', 'JJR']
- Verb Tags = ['VB', 'VBP', 'VBD', 'VBG', 'VBZ']
- Adverb Tags = ['RB', 'RBR', 'RBS']

Adverb and adjective tags do not have much significance in generating NoSQL queries. Only noun and verb tags are considered for the next steps of PoS tagging. Because verb & noun tags may indicate command and attributes or table name respectively. Algorithm 2 illustrates the process.

**B. Collection and Attribute Extraction**

Levenshtein distance (LD) algorithm is used in a specific solution to extract collections and attributes from natural

**Algorithm 2: Keeping Necessary Tags**

**Input:** W = All the words after removing stop words;  
**Output:** T = Necessary Tag's with appropriate word  
 $c = \text{CountWord}(W)$   
**for**  $c \in C$  **do**  
     $w = W[c]$  **TAGS** = Tagging(w)  
    **for**  $t \in \text{TAGS}$  **do**  
        **TAG** = EMPTY  
        **if**  $t \in \text{VERB}$  **then**  
            |  $\text{PUSH}(\text{TAG}, t)$   
        **end**  
        **if**  $t \in \text{NOUN}$  **then**  
            |  $\text{PUSH}(\text{TAG}, t)$   
        **end**  
    **end**  
     $\text{PUSH}(T, \text{TAG})$   
**end**  
**return** T

language queries. The approach starts by counting how many words in the list are similar to one another. Afterward, it compares every single similar word with every attribute from the WordNet by the LD and synonym list plays a crucial role in extracting attribute and collection names. This method keeps a list of words that are synonyms for each noun tag. Using WordNet, a list of noun tag synonyms is generated. The aim of making a synonym list is to find a specific collection and attribute from an input query. Every user formulates their query in a different way. They also use different words to describe the attribute or collection names. So, this approach checks synonyms of the words from the user query in the WordNet library. we give some analogies in Table II:

TABLE II. ANALOGY BETWEEN TEXT AND INTRUSION DETECTION WHEN APPLYING THE LD ALGORITHM

Text	Intrusion Detection
Find the name of all <b>student</b>	Collection( <b>all_student</b> )
What is the <b>accommodation</b> of student id 01	Attribute( <b>address</b> )
Find the <b>Fastname</b> of the students	Attribute( <b>name</b> )

In this paper, the LD algorithm works as a threshold. Sentence word is compared with the collection name if the value is greater than the threshold then it saves the collection and attributes with the appropriate name in a list. Finally, this approach gives an output figure of the match collection and attribute. The designed algorithm for collection and attribute extraction is described in Algorithm 3

Levenshtein Distance formula is used to measure the distance between the two strings  $a$  and  $b$  with length  $|a|$  and  $|b|$ , respectively.

$$LD_{a,b}(m, n) = \begin{cases} \max(m, n) \\ \min \begin{cases} LD_{a,b}(m-1, n) + 1 \\ LD_{a,b}(m, n-1) + 1 \\ LD_{a,b}(m-1, n-1) + 1_{(a_m \neq b_n)} \end{cases} \end{cases}$$

Here  $(a_m \neq b_n)$  is the indicator function that is equal to 0 when  $(a_m \neq b_n)$ , otherwise 1, and  $LD_{a,b}(m, n)$  is the distance

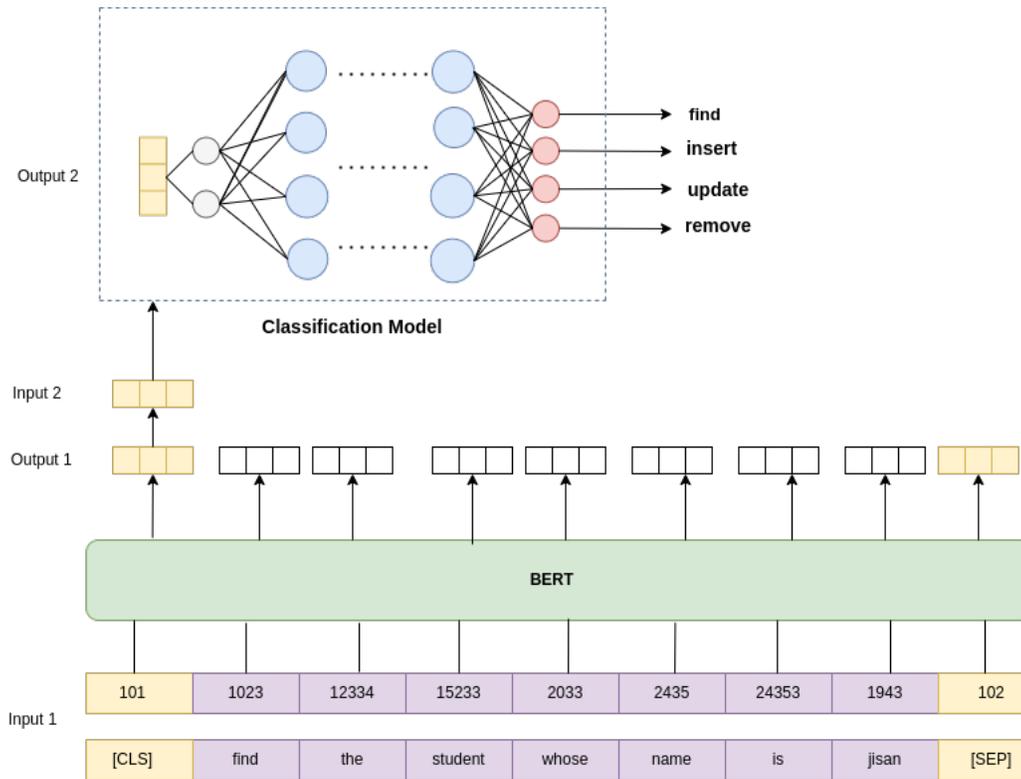


Fig. 2. Operation extraction using BERT model

**Algorithm 3: Attribute Extraction**

```

Input: W = List of Attributes from Database; C =
List of Collection name from Database; S = Set of
Similar Words
Output: A = Attributes Name; B = Table Name
t = CountWord(T)
for i ← 1 to t do
  for j ∈ S do
    LD - THRESHOLD = 1
    THRESHOLD = LD-Algorithm(S[j], W[j])
    if LD - THRESHOLD > THRESHOLD
      then
        PUSH (A[i], W[j])
        PUSH (B[i], C[i])
      end
    end
  end
end

```

between the first m characters of a and the first n characters of b.

*C. Operation Extraction*

Operation extraction is a particular solution that uses BERT Model to extract operations from natural language queries. In this approach, we use BERT Model for classifying the specific operation. In machine learning, classification is the set of categories that analysis belongs to the basis of a training set of

data containing (or instances) whose categorical membership is known [36]. A classification model tries to make some inferences from the observed data. To predict one or more outcomes from the dataset, provide one or more data as inputs to the categorization model.

In the dataset, BERT employs a novel technique known as Masked Language Model (MLM), in which it masks words in the sentence at random and then attempts to predict them. It doesn't use common sequence left-to-right or right-to-left language models. Instead, it uses the bidirectionally trained sequence with a deeper sense of language context and the model. The pre-train BERT applying two unsupervised tasks:

- Pre-training the BERT to understand language.
- Fine-tuning the BERT to learn specific task.

BERT depends on a Transformer (the self-attention mechanism to learn contextual relationships between words in a text). A simple Transformer consists of an encoder that reads text input and a decoder to generates a task prediction. Since the BERT model only requires the encoder part for generating a language representation model. There are two main models of BERT:

- **BERT base** has 12 transformer blocks, 768 hidden layers, 12 attention heads, and 110M parameters.
- **BERT large** has 24 transformer blocks, 1024 hidden layers, 16 attention heads, and 340M parameters.

In this paper, we used the BERT base model that has enough pre-trained data to help bridge the gap in data. The

model for operation extraction shows in Fig. 2. Given the input text, the Model that tokenizes the text using BERT tokenizer then generates the input masks with input IDs of the sentence. The input mask uses WordPiece [37] for tokenizing that splits the token like “going” to “go” and “ing.” It is mainly to cover a broad spectrum of Out-Of-Vocabulary (OOV) words. After tokenization, the output class goes as input in the classification model. we used a neural network for classification to get the highest accuracy. After classifying, we get the output of the operation. Here we work on four types of operations, in consideration- **FIND, INSERT, UPDATE, REMOVE**.

#### D. Build Syntax Tree & Generate Query

After Tokenization, collection, attribute, and operation are extracted from the sentence, we map the syntax tree with key-value pairs to build the query sequentially with the logical expression. If there are no logical expression in the sentence, it will be Nulled. Fig. 3 shows the syntax tree.

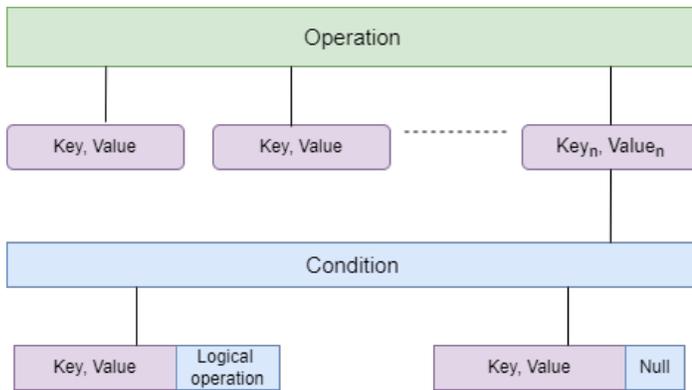


Fig. 3. Mapping syntax tree

Finally, we concatenate the whole step part-by-part and generate a NoSQL query. Fig. 4 shows the architecture of NoSQL query and given the output of the result.

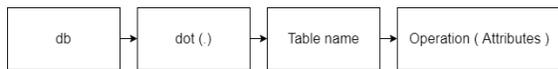


Fig. 4. Architecture of query generation

## IV. EXPERIMENTAL ANALYSIS AND RESULT

In this section, we evaluate our proposed model with the dataset. Firstly, we present the analysis of our dataset, then set up the evaluation. In the end, we compare our proposed model with the existing works and mention the differences, weak and strong points of our proposed model.

#### A. Dataset

We reshuffle the WIKISQL dataset for a better understanding of our model performance. WIKISQL is a massive crowd sourced dataset for creating NLIDB. The model is retrained periodically by reflecting the latest dataset. Our proposed model has used two types of data: (1) Natural Language

Query column which represent the natural language query and (2) Operations column. The description of the datasets is illustrated in Table III.

TABLE III. DESCRIPTION OF DATASET

Dataset	WIKISQL
Language	NLQ
Total number of cases	80,654
Length of the text (average)	61.09
Word count of the text (average)	11.66
Granularity of text description	line
Number of validation text	8,421
Number of test cases (total)	15,878
Number of train cases (total)	56,355

To avoid overfitting, we split the dataset into the training set and the testing set. we train our model on 70:30, 60:40, and 80:20 ratios and get the optimal result from the 80:20 ratio on our dataset. The data fields are the same among all splits. WikiSQL is a collection of hand-annotated SQL table, question, and query examples from Amazon Mechanical Turk crowd workers. It is orders of magnitude larger than current datasets, with 87000 samples as of this writing. The number of validation queries is 8,421. We build queries for the table and then ask crowd workers to paraphrase them. Each paraphrase is then double-checked by independent personnel to ensure that it does not alter the meaning of the original inquiry. We anticipate that making WikiSQL available will aid the community in developing the next generation of natural language interfaces (Fig. 5).

The Fig. 6 illustrates a blue histogram which shows the word and text distribution of dataset. It is hand-annotated semantic parsing dataset that contains logical and normal forms, respectively. In the dataset, the data is extracted from the web.

#### B. Text Pre-Processing

Text pre-processing is the first step of our proposed system. This step involves removing noise from our dataset. we apply several pre-processing steps to the data to convert words into numerical features. An example of tokenization is:

Input: ‘find the name of all student’  
Output: [‘find’, ‘the’, ‘name’, ‘of’, ‘all’, ‘student’]

#### C. Collection and Attribute Extraction from WordNet

We used NLTK WordNet to find find synonyms and antonyms of words. A WordNet is a lexical database that contains semantic relationships between words and their meanings. Our proposed model can successfully extract collection and attributes from WordNet library if there were any spelling errors occur or synonyms used. The bar diagram 5 shows how extract collection and attribute from WordNet using Levenshtein distance. For example:

Collection extraction: ‘all\_student’: [‘student’, ‘students’]  
Attribute extraction: ‘name’: [‘name’, ‘title’, ‘label’]

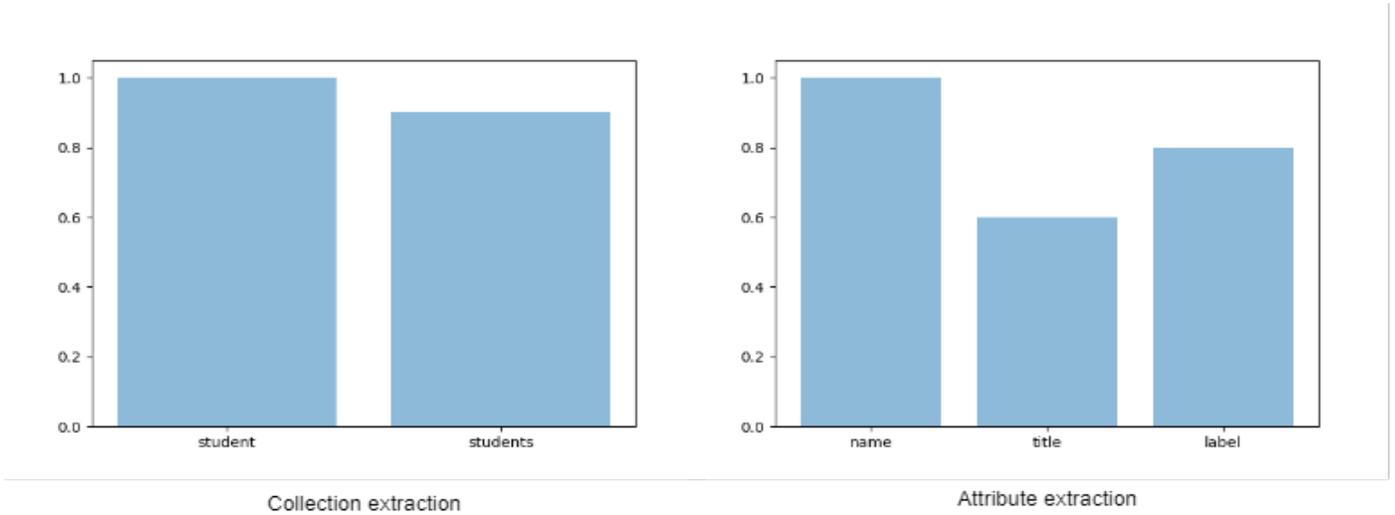


Fig. 5. Collection and attribute extraction

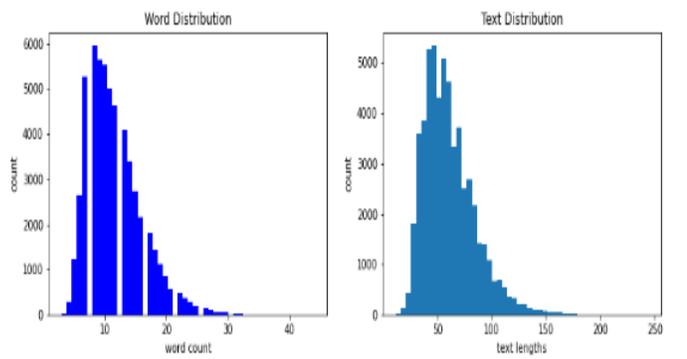


Fig. 6. Word and text distribution

#### D. BERT Tokenizer

In operation extraction our proposed system starts with BERT Tokenizer step. It gives sinusoidal positional encoding, the model itself learns the positional embedding during the training phase. Using the word-piece tokenizer concept that break some words into sub-words.

It helps many times to break unknown words into some known words and tokenize our text into tokens that correspond to BERT’s vocabulary. An example of BERT Tokenization is:

Input: ‘find the name of all student’  
Output 1: [101, 1023, 12334, 15233, 2033, 2435, 24353, 102]  
Output 2: [‘[CLS]’, ‘find’, ‘the’, ‘name’, ‘of’, ‘all’, ‘student’, ‘[SEP]’]

Output 1 is indices of the input tokens from the vocab file and output 2 is the reverse, a human-readable token of the input\_ids. Apart from the input tokens we also got 2 special tokens ‘[CLS]’ and ‘[SEP]’. BERT model is designed in such a way that the sentence has to start with the [CLS] token and end with the [SEP] token.

#### E. Split Data for Training and Testing

The training phase is the first step for the BERT Model. This model is a transformer design based on an encoder stack. We trained the WIKISQL dataset using this model. The Model uses the Semi-Supervised Learning approach for translating natural language query into operation. The training sub-dataset contains all of the features required to turn a natural language query into an operational query. To partition the WIKISQL dataset into two sub-datasets, we use the scikit-learn library’s “train test split” method. The suggested system is built using the dataset’s training sub-dataset. The training dataset is a fraction (80%) of the whole data set. The rest (20%) is considered as test data. This information is imported as a.csv file. Table IV shows a portion of the training data set.

TABLE IV. A SAMPLE OF TRAINING DATASET

Line No.	Natural Language Query	Operations
1	What’s Dorain Anneck’s pick number?	find
2	Find the student whose name is x.	find
3	Insert the arrival time of greenbat.	insert
4	Put the status of the trains at location Museum	insert
5	Update the record for september 15, 1985.	update
6	Re-equip the student	update
7	Remove the brighton cast for jerry cruncher	remove
8	Delete the all student	remove

#### F. Model Building

BERT is an architecture that uses a transformer encoder to process each token of input text in the context of all other tokens. After splitting the dataset, we start with the pre-trained BERT Model to classify the find, insert, update and remove operations. In our model we use 12 layers of Transformer encoder. After run the operation we get two variables: First

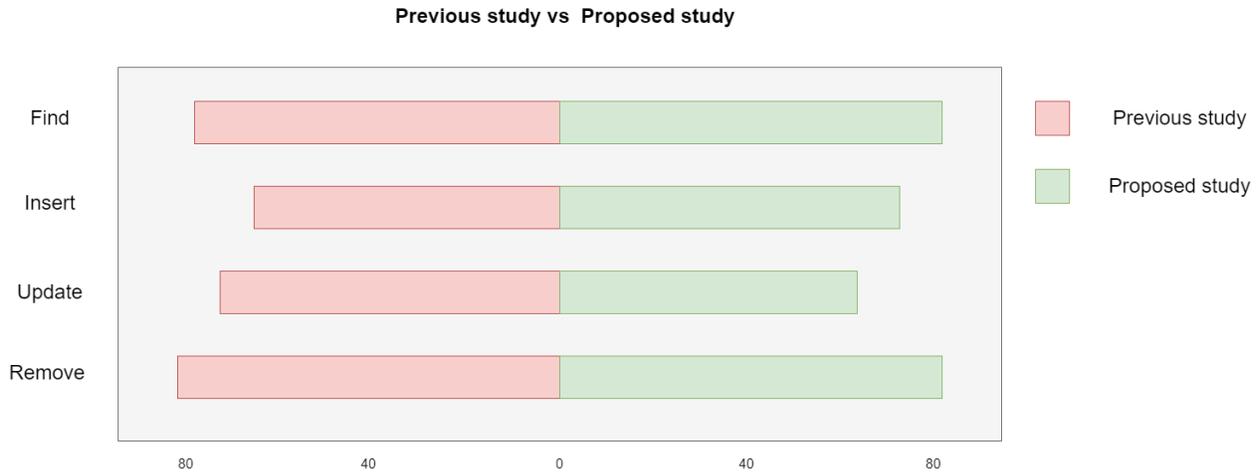


Fig. 7. Comparison of four types operations

variable contains the embedding vectors of all of the tokens in a sequence and second variable contains the embedding vector of [CLS] token. We then pass the variable into a linear layer with ReLU activation function. We have a vector of size 4 at the end of the linear layer, each of which corresponds to a category of our labels (find, insert, update, and remove). We use Adam as the optimizer and train the model for 10 epochs. Because we're dealing with multi-class classification, we'll need to use categorical cross entropy as our loss function. Fig. 8 depicted the operation.

For example:

Input: 'find the name of all student'  
Output: 'find'

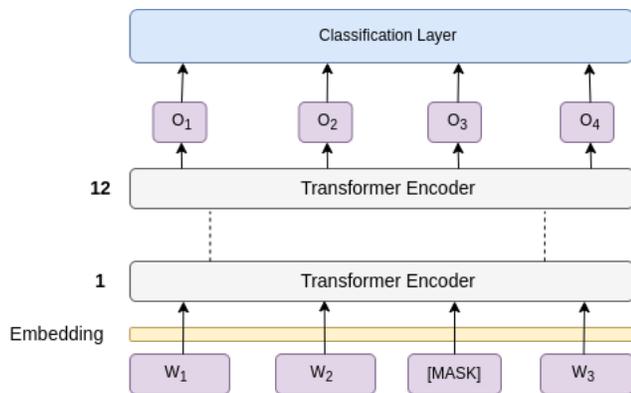


Fig. 8. Model building

The model enhances the accuracy rate for classification than the previous model. For the classification task, the model can classify 81.45% average class detection from previous research. One of the reasons is BERT uses a pre-trained model which is based on transfer learning. It can tune the data on a specific NoSQL language. Fig. 7 illustrates the accuracy rate of four types of operations separately.

### G. Model Accuracy

Accuracy evaluates how well our model forecasts compare them with the original values. With a low rigor yet a high blunder, the model would make huge mistakes in the data. Both blunder and rigor lowness indicates that with most data, the model produces smaller errors. However, it produces huge mistakes in some systems if they are both high. The ideal scenario of any model would be high rigor and little blunder. Fig. 9 illustrate the accuracy of the proposed model.

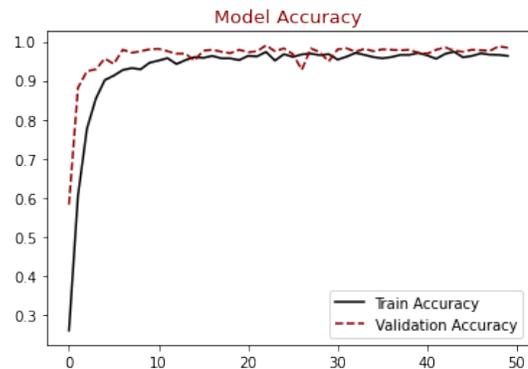


Fig. 9. Model accuracy

### H. Model Loss

Loss is the total of our model errors. It evaluates how well our model does (or how badly it does). When there are a lot of mistakes, the loss is high and the model doesn't work properly. The better our model works, the lower it is. However, the greatest conclusion we can make from it is whether the loss is big or low. If we plot losses over time, we can evaluate if and how quickly our model is learning. This is because the loss function is utilized by the model for learning. This takes the shape of approaches like gradient descent, which modify parameters of the model using information on the loss outcome. Fig. 10 illustrate the loss of the proposed model.

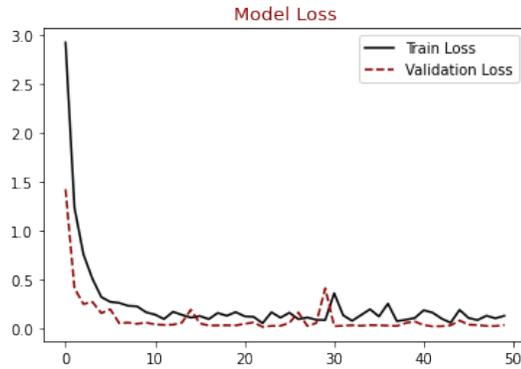


Fig. 10. Model loss

### I. Output

In output, we get the collection and attribute name, such as **all\_student** and **name**. From the operation extraction, get the **find** operation, then concatenate all the extractions output part-by-part to generate a NoSQL query. For example:

```
db.all_student.find({}, {'name':True})
```

We have classified the wrong output into two categories: (a) sometimes, the query contained incomplete logical expression in condition part (b) the query is incorrect. Analysis of the conversion results reveals the following:

- Observing all the NoSQL output, we can notice suggested model can work with natural language queries of different lengths. After a successful NoSQL query output, the number of input and output tokens might be distinct. The accuracy of the proposed model did not depend on the length of the query.
- The BERT Model successfully predicts the operation using a pre-trained model. It also tunes the NoSQL command from a distinctive size of input text.
- The BERT model can process a large amount of data. The WIKISQL dataset covered different types of query statements. So there is no problem for the BERT model to work with the WIKISQL dataset.
- The Bert model understands the semantic relationship between natural language and NoSQL queries. As a result, the decoder output is logically correct for the maximum query.
- The model can generate “contextualized” word embeddings but it is compute-intensive at inference time and need to calculate compute vectors every time.
- In collection and attribute extraction, we use the Levenshtein Distance algorithm. The algorithm can extract attributes from natural language queries furthermore check the spelling error. The run time complexity of this algorithm is lower than  $O(n^2)$ .

Test results show in Table V that have been translated into the NoSQL syntax. The test data contains the natural

language query as well as appropriate. The output contains each converted NoSQL query with original query and test query, along with the percentage of converted NoSQL query.

### J. Evaluation Setup

In this dissertation, we evaluate the result on our dataset that have three notation to evaluate the query synthesis accuracy.

- **Normal form accuracy** is the form of a NoSQL query that has no attribute. We analyze the synthesized NoSQL query with the ground truth to verify whether they match each other.
- **Logical form accuracy** is the accuracy of a NoSQL query that has attributes or any logical expression of the query.
- **Query match** is the comparison accuracy with the original query match for find, insert, update and remove operations query. We use a canonical representation of the synthesized NoSQL query and the ground truth to determine whether two NoSQL queries are identical.

We also find out the F1 score for operation extraction that measures the precision and recall value. Finally, we present the comparison of our model with previous work on NoSQL conversion tasks. The implementation of our model using python [38].

The F1-score measures the accuracy of the operation (find, insert, update, remove) by applying the precision and recall values of the test. This test looks at whether the system can process the sentences entered by the user so that it can be measured the operation accurately with the F1-score method. Table VI shows the accuracy values. The equation of the F1-score, precision, recall, and accuracy have given below:

- **Precision:** It is the true positive relevance rate that defined as the ratio  $\frac{tp}{tp+fp}$ , where  $fp$  indicates the number of false positives;
- **Recall:** It is the true positive rate that defined as the ratio  $\frac{tp}{tp+fn}$ , where  $tp$  and  $fn$  are the number of true positives and false negatives, respectively;
- **F1-score:** F1-score is a function of Precision and Recall that is the harmonic mean between Precision and Recall, defined the ratio as  $\frac{2*(precision*recall)}{precision+recall}$ ;

Next, we find out the accuracy of normal and logical forms. Let  $X$  is the total number of queries in our dataset and  $X_{ex}$  is the execution query. we evaluate the every clause (find, insert, update and remove) query using accuracy metric for normal form  $Acc_{nf} = \frac{X_{ex}}{X}$  and for logical form  $Acc_{lf} = \frac{X_{ex}}{X}$ . Table VII shows the accuracy of normal and logical queries. After that the overall result is evaluated by the BLEU (Bilingual Evaluation Understudy) that was developed to evaluate the machine translation system.

### K. Result

The article presents an efficient approach to transform the natural language query into a NoSQL query effectively.

TABLE V. THE ACCURACY FOR CONVERTING NATURAL LANGUAGE INTO NON STRUCTURED QUERY LANGUAGE

Input Text	Original query	Test query	Accuracy(%)
Find all the students	db.all_student.find()	db.all_student.find()	100
What is the name of all student?	db.all_student.find({name:True})	db.all_student.find({name:True})	100
Find the student whose age greater than 70	db.all_student.find({ age: { \$gt: 70 } })	db.all_student.find({ age: { 70 } })	75.0
Insert the student whose name is x	db.all_student.insert ({name:'x'})	db.all_student.insert ({name:'x'})	100
Insert student whose name is x, age 22	db.all_student.insert ({name:'x',{age:22}})	db.all_student.insert ({name:'x',{age:22}})	83.33
Update the name y who is x in student table	db.all_student.update ({name:'x'},{set:{name:'y'}})	db.all_student.update ({name:'x'},{set:{name:'y'}})	100
Update name z and age 40, whoes name is x	db.all_student.update ({name:'x'},{set:{name:'y'}})	db.all_student.update ({name:'x'},{set:{name:'x'}})	65.5
Remove all the students	db.all_student.drop()	db.all_student.drop()	100
Delete student whose name is x and age 20	db.all_student.removeMany ({name:'x', age: 20})	db.all_student.remove ({name:'x', age: 20})	75.0

This model achieves a competitive result on our dataset. The following tables represent the experiment result of each classifier.

TABLE VI. EXPERIMENTAL RESULTS OF EACH CLASSIFIER

F1-score	0.808
Precision	0.892
Recall	0.74

Bilingual Evaluation Understudy (BLEU) is a score for comparing a candidate translation of the NoSQL query to one or more reference translations. To predict the accuracy of automatic machine translation systems, Kishore Papineni, et al. [39] proposed this score in 2002. We used the BLEU score to determine the output.

BLEU is not entirely effective but offers several interesting benefits like quick, easy to calculate, language-independent, highly interactive with human interpretation, and widely used.

$$P = \frac{m}{w_t} \quad (1)$$

where,  $m$  is the estimate of tokens from the candidate source code that are found in the reference text, and  $w_t$  is the total estimate of words in the candidate query. The accuracy is calculated using the equation 2.

$$Accuracy = P \times 100\% \quad (2)$$

The performance analysis of our model is given in Table VII.

TABLE VII. PERFORMANCE ANALYSIS OF OUR MODEL.  $Acc_{nf}$  AND  $Acc_{lf}$  INDICATE THE NORMAL FORM AND LOGICAL FORM QUERY ACCURACY, AND  $Acc_{qm}$  INDICATES THE ACCURACY OF QUERY MATCH

Operation clause	$Acc_{nf}(\%)$	$Acc_{lf}(\%)$
Find	100	87.5
Insert	-	91.67
Update	-	82.75
Remove	100	75.0

Accounting to Concepts Identification errors and domain dictionary errors, the average accuracy achieved by our system is 88.76% respectively. We define the error rate as:

$$ErrorRate = (100 - AccuracyRate)\% \quad (3)$$

TABLE VIII. ANALOGY OF DIFFERENT TYPES OF MODEL

Model	Accuracy (%)	Error Rate (%)
Encoder-Decoder Model	71.5	28.5
REINFORCE-algorithm Model	84.2	15.8
Proposed Model	88.76	11.24

Table VIII represents BLEU portion of efficiency for forecasting correct NoSQL query. Using the WikiSQL reshape dataset the proposed model is passed down for comparing with the existing other models. Fig. 11 illustrates the three models' estimated efficiency and error rates. It demonstrates the accuracy of other measure rates of converting the natural language query into the non-structured query language (that scored 88.76%) is better or at least competitive than the earlier results.

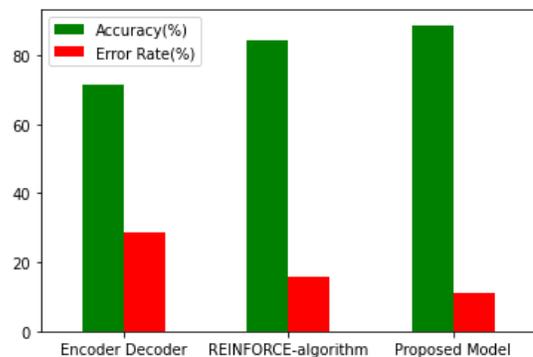


Fig. 11. Performance factor between previous and our proposed model

## V. CONCLUSION

In the age of digitalization, internet users have been increasing continuously. So a large amount of data needs to be stored in a database. Relational databases faced some challenges in search engines and social networking services. Here, the NoSQL database helps maintain a broad range of hierarchical data models. The proposed model deals with the NLP

to non-relational query conversion. Initially, preprocessing the text (English) by NLTK, then used LD algorithm for collection, attribute extraction and BERT model for operation extraction and finally, query generation. Our model can generate queries for Find, Insert, Update, Remove clause with an average accuracy of 88.76%. In the future, we intend to improve more complex NoSQL queries such as logical function queries, using other incentive mechanisms for better performance.

#### ACKNOWLEDGMENT OF FUNDING

This work was supported by the UGC Jagannath University Research Branch, Dhaka, Bangladesh, Under JnU/research/rp/2020-21/science/44.

#### REFERENCES

- [1] J. Han, E. Haihong, G. Le, and J. Du, "Survey on nosql database," in 2011 6th international conference on pervasive computing and applications. IEEE, 2011, pp. 363–366.
- [2] A. Nayak, A. Poriya, and D. Poojary, "Type of nosql databases and its comparison with relational databases," International Journal of Applied Information Systems, vol. 5, no. 4, pp. 16–19, 2013.
- [3] R. S. Al Mahruqi, "Migrating web applications from sql to nosql databases," Ph.D. dissertation, Queen's University (Canada), 2020.
- [4] S. Batra, C. Tyagi, "Comparative Analysis of Relational And Graph Databases", IJSCE, vol.2(2), pp. 509-512, 2012.
- [5] R. Alexander, P. Rukshan, and S. Mahesan, "Natural language web interface for database (nlwdb)," arXiv preprint arXiv:1308.3830, 2013.
- [6] Z. Wei-ping, L. Ming-xin and C. Huan, "Using MongoDB to implement textbook management system instead of MySQL", IEEE-ICCSN, 2011, pp. 303-305.
- [7] P. Chen, C. Xhang, "Data-intensive applications, challenges, techniques and technologies: A survey on Big Data", Information Sciences, Elsevier, vol.275, pp.314–347, 2014.
- [8] B. Jose, S. Abraham, Unstructured Data Mining for Customer Relationship Management: A Survey, International Journal of Management, Technology And Engineering 8, Issue 7. ISSN NO (2018) 2249–7455.
- [9] O. Ferschke, J. Daxenberger, and I. Gurevych, "A survey of nlp methods and resources for analyzing the collaborative writing process in wikipedia," in The People's Web Meets NLP. Springer, 2013, pp. 121–160.
- [10] Garrido-Muñoz, A. Montejó-Rázquez, F. Martínez-Santiago, and L. A. Ureña-López, "A survey on bias in deep nlp," Applied Sciences, vol. 11, no. 7, p. 3184, 2021.
- [11] S. Srivastava, A. Shukla, and R. Tiwari, "Machine translation: from statistical to modern deep-learning practices," arXiv preprint arXiv:1812.04238, 2018.
- [12] Kłosowski, P. (2018). Deep learning for natural language processing and language modelling. In 2018 Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA), September 2018, pp. 223-228. IEEE. 10.23919/SPA.2018.8563389
- [13] U. K. Acharjee, M. Arefin, K. M. Hossen, M. N. Uddin, M. A. Uddin, and L. Islam, "Sequence-to-sequence learning-based conversion of pseudo-code to source code using neural translation approach," IEEE Access, vol. 10, pp. 26 730–26 742, 2022.
- [14] B. Jose and S. Abraham, "Intelligent processing of unstructured textual data in document based nosql databases," Materials Today: Proceedings, 2021.
- [15] N. Yaghmazadeh, X. Wang, and I. Dillig, "Automated migration of hierarchical data to relational tables using programming-by-example," Proceedings of the VLDB Endowment, vol. 11, no. 5, pp. 580–593, 2018.
- [16] S. Zhang, Y. Hu, and G. Bian, "Research on string similarity algorithm based on levenshtein distance," in 2017 IEEE 2nd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC). IEEE, 2017, pp. 2247–2251.
- [17] T. Ho, S.-R. Oh, and H. Kim, "A parallel approximate string matching under levenshtein distance on graphics processing units using warp-shuffle operations," PLoS one, vol. 12, no. 10, p. e0186251, 2017.
- [18] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," CoRR, vol. abs/1810.04805, 2018.
- [19] Woods, W. A., 1973. Progress in natural language understanding: An application to LUNAR geology. AFIPS Natl. Computer. Conj. Expo.. Conference Proc. 42, 441-450.
- [20] M. Arefin, K. M. Hossen, and M. N. Uddin, "Natural language query to sql conversion using machine learning approach," in 2021 3rd International Conference on Sustainable Technologies for Industry 4.0 (STI). IEEE, 2021, pp. 1–6.
- [21] B. Mallikarjun, K. Annapoorneshwari, M. Yadav, L. R. Rakesh, and S. Suhaas, "Intelligent automated text processing system-an nlp based approach," in 2020 5th International Conference on Communication and Electronics Systems (IC- CES). IEEE, 2020, pp. 1026–1030.
- [22] B. Jose and S. Abraham, "Intelligent processing of unstructured textual data in document based nosql databases," Materials Today: Proceedings, 2021.
- [23] F. Abdelhedi, A. A. Brahim, and G. Zurfluh, "Ocl constraints checking on nosql systems through an mda-based approach," International Journal of Data Warehousing and Mining (IJDWM), vol. 17, no. 1, pp. 1–14, 2021.
- [24] M. T. Majeed, M. Ahmad, and M. Khalid, "Automated xquery generation for nosql," in 2016 SIXTH International Conference on Innovative Computing Technology (INTECH). IEEE, 2016, pp. 507–512.
- [25] S. Mondal, P. Mukherjee, B. Chakraborty, and R. Bashar, "Natural language query to nosql generation using query-response model," in 2019 International Conference on Machine Learning and Data Engineering (iCMLDE). IEEE, 2019, pp. 85–90.
- [26] T. Pradeep and P. C. Rafeeqe and Murali, Reena, Natural Language To NoSQL Query Conversion using Deep Learning (August 13, 2019). In proceedings of the International Conference on Systems, Energy & Environment (ICSEE) 2019, GCE Kannur, Kerala, July 2019, Available at SSRN: <https://ssrn.com/abstract=3436631> or <http://dx.doi.org/10.2139/ssrn.3436631>
- [27] S. Blank, F. Wilhelm, H.-P. Zorn, and A. Rettinger, "Querying nosql with deeplearning to answer natural language questions," in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, no. 01, 2019, pp. 9416–9421.
- [28] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," Advances in neural information processing systems, vol. 27, 2014.
- [29] M. D. Gadekar, B. M. Jadhav, A. S. Shaikh, and R. B. Kokare, "Natural language (english) to mongodb interface," International Journal of Advanced Research in 9 Computer Engineering & Technology (IJARCT), vol. 4, no. 3, 2015.
- [30] P. Anand and Z. Farooqui, "Rule based domain specific semantic analysis for natural language interface for database," International Journal of Computer Applications, vol. 164, no. 11, 2017.
- [31] T. Mahmud, K. A. Hasan, M. Ahmed, and T. H. C. Chak, "A rule based approach for nlp based query processing," in 2015 2nd International Conference on Electrical Information and Communication Technologies (EICT), pp. 78–82, IEEE, 2015.
- [32] X. Xu, C. Liu, and D. Song, "Sqlnet: Generating structured queries from natural language without reinforcement learning," arXiv preprint arXiv:1711.04436, 2017.
- [33] J. Bornholt, E. Torlak, D. Grossman, and L. Ceze, "Optimizing synthesis with metasketches," in Proceedings of the 43rd Annual ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages, 2016, pp. 775–788.
- [34] V. Zhong, C. Xiong, and R. Socher, "Seq2sql: Generating structured queries from natural language using reinforcement learning," arXiv preprint arXiv:1709.00103, 2017.
- [35] K. Guu, P. Pasupat, E. Z. Liu, and P. Liang, "From language to programs: Bridging reinforcement learning and maximum marginal likelihood," arXiv preprint arXiv:1704.07926, 2017.
- [36] G. B. Boullanger and M. Dumonal, "Search like a human: Neural machine translation for database search," Technical report, Tech. Rep., 2019.

- [37] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey et al., "Google's neural machine translation system: Bridging the gap between human and machine translation," arXiv preprint arXiv:1609.08144, 2016.
- [38] M. K. Chakravarthy and S. Gowri, "Interfacing advanced nosql database with python for internet of things and big data analytics," Materials Today: Proceedings, 2021.
- [39] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in Proceedings of the 40th annual meeting on association for computational linguistics, pp. 311–318, Association for Computational Linguistics, 2002.

# Privacy-Preserving and Trustless Verifiable Fairness Audit of Machine Learning Models

Gui Tang<sup>1</sup>, Wuzheng Tan<sup>2</sup>, Mei Cai<sup>3</sup>,  
College of Cyber Security, Jinan University, Guangzhou, China<sup>1,2</sup>  
Jinan University Library, Guangzhou, China<sup>3</sup>

**Abstract**—In the big data era, machine learning has developed prominently and is widely used in real-world systems. Yet, machine learning raises fairness concerns, which incurs discrimination against groups determined by sensitive attributes such as gender and race. Many researchers have focused on developing fairness audit technique of machine learning model that enable users to protect themselves from discrimination. Existing solutions, however, rely on additional external trust assumptions, either on third-party entities or external components, that significantly lower the security. In this study, we propose a trustless verifiable fairness audit framework that assesses the fairness of ML algorithms while addressing potential security issues such as data privacy, model secrecy, and trustworthiness. With succinctness and non-interactive of zero knowledge proof, our framework not only guarantees audit integrity, but also clearly enhance security, enabling fair ML models to be publicly auditable and any client to verify audit results without extra trust assumption. Our evaluation on various machine learning models and real-world datasets shows that our framework achieves practical performance.

**Keywords**—Security and privacy; machine learning; fairness; cryptography; zero knowledge proof

## I. INTRODUCTION

Machine Learning has seen great success in decision-making and decision-support tasks in recent years [1] [2] [3], being deployed in various applications and products in practice, such as loans and hiring decisions. However, concerns are rising that algorithms amplify bias and discrimination from the training data, and fairness is becoming an essential metric for evaluating machine learning models [4]–[8]. Consequently, fairness has become a roadblock to widespread machine learning applications. To address this formally, many works towards considering how algorithm fairness can be assessed by proposing various measures and how discrimination in machine learning systems can be mitigated by pre-processing [9], [10], inter-processing [11], [12], and post-processing methods [5] [13].

In practice, there is a need for guarantees that the result of fairness audit are correctly calculated with respect to specific fairness metrics, which is referred to as the *audit integrity* of fairness. One of the basic ideas to ensure that users are protected from discrimination is to ensure the integrity of the audit. In order to get a fair model, the server usually requires the user's sensitive data, such as gender and race, to train the machine learning model. However, this requirement is often contrary to the interests of the user. First, users are usually reluctant to share their data, even if it is a reasonable aim,

because it would expand their exposure to privacy risks. In addition, the collection of sensitive user data is subject to legal restrictions. For example, the EU's General Data Protection Regulation (GDPR) highlights the minimal prerequisites for collecting sensitive data [14]. If the model itself is not a secret, anyone can potentially run tests on it to establish its purported fairness without exposing its data. However, this approach may be contrary to the benefits of the model owner due to intellectual property. Although there are fair learning approaches [15], [16], training fair models without the sensitive data have been proposed, it is still required to have the sensitive data for assessing the fairness of the trained model [17]. We call this problem as *sensitive data availability*.

To overcome sensitive data availability issues in providing audit integrity, Veale and Binns [14] introduce a trusted third party with sensitive data to certify the fairness of a machine learning model. Although this model works well, it requires a strong trust relationship between the third party and the model owner. Either the third party has access to the ML model, or the model owner has access to the sensitive data, which may be against their interests. To audit the model publicly while protecting sensitive data's privacy and keeping the model confidentiality, Kilbertus et al. [18] and Segal et al. [19] proposed to utilize multi-party computation (MPC) approach. Those approaches enable a public fairness audit under the assumption of a semi-honest security model and are extended by Pentylala et al. [20] to a malicious security model. Park et al. [17] propose a framework to enable secure fairness audit by leveraging confidential computing based on hardware enclave under the malicious security model.

**The problem.** While previous work have been work well, existing solutions still suffer from extra trust assumption. This is problematic for two reasons. First, the additional trust assumptions mean the third party determines the audit integrity. Second, relying on a third party can lead to single-point failures. Specifically, the MPC-based approach assumes that the server and the third party are running all required steps in a protocol. Moreover, the hardware-based approach introduces additional hardware security assumptions and also suffers from hardware vulnerabilities [21] [22]. Furthermore, when we are in a situation where we want to audit the model used for several different domains, we need to establish credible relationships with more third parties or hardware enclaves.

To get secure and robust service, we need a much more robust security guarantee: each party only trusts itself. This raises our question: *Can we design a framework for auditing the fairness of machine learning models under no trusted party existing scenario? Or can we guarantee security for audit*

<sup>2</sup>Corresponding Author.

integrity without external trust assumptions? For example, we want to support fairness audits as a service in the model market to achieve fairness integrity.

We answer the question above positively in this paper by proposing a fair audit framework, which enables a publicly verifiable fairness audit of the ML model without disclosing model parameters and guarantees audit integrity of the fair audit. The main idea is to leverage the progress of zero knowledge succinct non-interactive arguments of knowledge (zk-SNARKs) [23]–[29] recently. A zk-SNARK enables the third party to efficiently convince the verifier that the computation of fairness audit is correctly calculated. We solve the critical challenge of adapting zk-SNARK to this work under the malicious threat model. In summary, the contributions of this work are:

- We provide a generic framework to audit the fairness of machine learning model under the trustless condition. We can support generic machine learning models with arbitrary fairness metrics.
- We formally define security requirements and instantiate the framework described above. We have solved performance challenging problems in the instantiation process.
- We implement our framework and evaluate its performance on several real-world datasets. The experimental results show that our framework achieves practical performance.

## II. LITERATURE REVIEW

### A. Fairness Audit

Despite training a machine learning model is a fundamental problem, bringing the model to reality is also important. A fundamental question is how to ensure that the model used is non-discriminatory. There is a line of work to discuss this problem. Veale and Binns [14] introduced highly trusted third parties selectively storing data and performing discrimination auditing to achieve fairness in machine learning. However, they assume the modeler must disclose their model to a third party or trust it in order to obtain the model prediction on test data, which may be incompatible with modeler's intellectual property. To resolve these problems, other privacy-preserving approaches such as multi-party computation or trusted execute environment can be applied. Kibertus et al. [18] and Segal et al. [19] proposed privacy-preserving fair certification and inference of ML model that protect sensitive attributes and model confidentiality by using MPC. However, they assume that two honest-but-curious server and require high communication. The following work PrivFair [20] extend their security model to active security threat models in 2- or 3-server setups. Park et al. [17] provided a generic fairness audit framework that relies on hardware enclaves and explores more potential threats and attacks in the fairness certification process. Although their approach has a small computational overhead, their require additional hardware and trustworthiness to TEE, which is not our goal. And TEE also face many unknown vulnerability [21]. All of these work require additional trusts, and does not provide public verifiable. Their computation integrity rely on their trust on third party. In this study, we explore a publicly

verifiable security audit protocol based on zero knowledge proof with lower level of trust. Also [19] and [17] explore the auditing dataset are publicly known during model training, which makes the model certification harder by allowing the modeler can adaptive training their model on the audit data. We are also using this approach to improve the reliability and robustness of fairness audits, which is seen as a promising direction for fairness certification.

### B. Zero Knowledge Proof

Zero knowledge proofs were introduced by Goldwasser [30] and generic constructions based on probabilistically checkable proofs were proposed in the seminal works of Kilian [31] and Micali [32]. In recent years there has been significant progress in efficient ZKP protocols and systems. A radically different approach in zero-knowledge proof, categorized by their underlying techniques and assumptions, there are pairing-based schemes [24], [25], [27], discrete- log-based schemes [33], interactive-proof [34], [35], interactive oracle proofs (IOP) [26], [36], and so on. They provide different trade-offs between prover runtime, proof size and verifier runtime and so on. Please refer to [37] for more details on the performance and comparisons of different ZKP schemes. Zero knowledge proof has been widely used in blockchains and cryptocurrencies to achieve privacy [38] and scalability. More recently, it also found new applications in zero-knowledge machine learning [39], [40], zero-knowledge middlebox [41], and so on.

## III. PRELIMINARIES

In this section, we introduce the fairness notions in machine learning and the cryptographic primitives used in our framework.

**Notions.** We use  $\lambda$  to denote the security parameter. Let  $[n]$  denote the set  $\{0, 1, \dots, n - 1\}$ ; a vector be denoted by a boldface letter, e.g.,  $\mathbf{x}$ . And  $x \leftarrow \mathcal{X}$  denote that  $x$  is sampled from a distribution  $\mathcal{X}$ .

### A. Fairness Notions in Machine Learning

There is plenty of fairness definitions [42] such as group fairness, causal discrimination, and counterfactual fairness. In this study, we mainly focus on statistical fairness definitions that require protected data, such as demographic parity (or statistical parity) [43], equalized odds [5], equal of opportunity [5], and disparate impact [44]. Demographic parity means that both protected and unprotected groups have an equal probability of being assigned to the positive predicted class. Equalized odds enforces both equal bias and equal accuracy in all demographics. Equal of opportunity is a relaxation of equalized odds, which only focus the positive predication outcome. Disparate impact implies that the decision outcomes disproportionately benefit or hurt members of certain sensitive attribute value groups.

Let  $M$  be a trained machine learning model for a classification task. Suppose possible inputs  $\mathcal{X}$ , sensitive or protected attribute  $\mathcal{G}$  (relevant for fairness, e.g., ethnic or sex), the true class label  $y$  and the predication  $\hat{y} = M(x, g)$ , where  $x \in \mathcal{X}, g \in \mathcal{G}$ . And we use tuple  $(M, \mathcal{X}, \mathcal{Y}, \mathcal{G})$  represent the audit sample  $D$ . Consider  $g = 0$  designates the unprotected

group and  $g = 1$  designates the protected group. We recall these fairness definition below.

- Demographic parity (DP):  $P(\hat{y} = 1|g = 0) = P(\hat{y} = 1|g = 1)$
- Equalized odds (EO):  $P(\hat{y} = 1|y, g = 0) = P(\hat{y} = 1|y, g = 1), \forall y \in \mathcal{Y}$
- Equal opportunity:  $P(\hat{y} = 1|y = 1, g = 0) = P(\hat{y} = 1|y = 1, g = 1)$
- Disparate impact (DI):  $P(\hat{y} = 1|g) \neq P(\hat{y} = 1), \forall g \in \mathcal{G}$

The fairness of the ML model is assessed by means of the empirical fairness gap, as described in [19]. The fairness notion must be expressed in the formulation in order to audit ML models. Without loss of generality, we can consider the demographic parity:

$$l_{g,y}(M) = \mathbb{E}_{(x,g',y')} [\mathbb{I}\{M(x) = y\} | g' = g]$$

where  $\mathbb{I}$  is an indicator function. Then define the estimated group risk as  $l_{g,y}(M, T) = \frac{1}{m} \sum_{i=1}^{m_g} \mathbb{I}\{M(x_i) = y_i \wedge g_i = g\}$ , where  $T = \{(x_1, g_1, y_1), \dots, (x_m, g_m, y_m)\}$  is independent sample set and  $m_g$  is the number of samples in  $T$  from group  $g$ . Define the fairness gap of ML model as

$$\max_{g_0, g_1 \in \mathcal{G}, y \in \mathcal{Y}} |l_{g_0,y}(M) - l_{g_1,y}(M)|$$

Also define empirical fairness gap (EFG) using the empirical approximation as follows:

$$EFG = \max_{g_0, g_1 \in \mathcal{G}, y \in \mathcal{Y}} |l_{g_0,y}(M, T) - l_{g_1,y}(M, T)|$$

We call model  $M$   $(\epsilon, \delta)$ -fair on  $(\mathcal{G}, \mathcal{T})$  with respect to a fairness measurement if:

$$\Pr \left[ \max_{g_0, g_1 \in \mathcal{G}, y \in \mathcal{Y}} |l_{g_0,y}(M) - l_{g_1,y}(M)| > \epsilon \right] \leq \delta$$

The EFG can naturally extend to the other fairness notion and yield the corresponding  $(\epsilon, \delta)$ -fairness definitions.

## B. Cryptography Primitives

**Bilinear Groups.** A bilinear group is given by a description  $GK = (p, \mathbb{G}_1, \mathbb{G}_2, \mathbb{G}_T)$  such that

- $\mathbb{G}_1, \mathbb{G}_2$  are cyclic groups of prime order  $p$  and generators are  $g \in \mathbb{G}_1$  and  $h \in \mathbb{G}_2$
- Bilinear map  $e : \mathbb{G}_1 \times \mathbb{G}_2 \rightarrow \mathbb{G}_T$ , that is,  $e(g^a, h^b) = e(g, h)^{ab}$ , where  $a, b \in \mathbb{Z}_p$
- $e(g, h)$  generates  $\mathbb{G}_T$

**Commitment Scheme.** A commitment scheme allows a committee to commit a secret value and later open the commitment and reveal the value to the verifier. We recall the commitment scheme definition.

*Definition 1:* A commitment scheme is a tuple of algorithms  $\text{Com} = (\text{Setup}, \text{Commit}, \text{VerifyCommit})$  that works as follows.

- $\text{Setup}(1^\lambda) \rightarrow \text{ck}$  takes as input the security parameter  $\lambda$  and outputs a commitment key  $\text{ck}$ .

- $\text{Commit}(\text{ck}, m) \rightarrow (c, o)$  takes as input the commitment key  $\text{ck}$  and a secret value  $m$ , and output a commitment  $c_m$  and an opening  $o$ .
- $\text{VerCom}(\text{ck}, c_m, m, o) \rightarrow b$  takes as input a commitment  $c_m$ , a value  $m$  and an opening  $o$ , and output accept ( $b = 1$ ) or reject ( $b = 0$ ).

The commitment scheme is required to be both binding and hiding. In the study, we will be using Pedersen-like commitment scheme [27] which is statistically hiding and computationally binding under suit assumptions.

**zkSNARKs.** Zero knowledge proof enables a *prover* to prove to a *verifier* the result  $y$  of a computation  $\mathcal{C}$  satisfying  $y = C(x, w)$ , where  $x$  is public input and  $w$  is secret witness of prover. The popular zero knowledge proof notions used in practice are zero-knowledge succinct non-interactive arguments of knowledge (zkSNARKs, for short) [23]–[25]. Here we recall the definition of zkSNARKs as follows:

*Definition 2:* A zk-SNARK for a relation  $R$  is a tuple of algorithms  $\Sigma = (\text{KeyGen}, \text{Prove}, \text{Verify}, \text{Sim})$  as follows:

- $\text{Setup}(1^\lambda, R) \rightarrow \text{CRS} = ((ek, vk), \tau)$ : The setup algorithm takes a relation  $R \in R_\lambda$  and security parameter as input, and returns a common reference string CRS and a simulation trapdoor  $\tau$ .
- $\text{Prove}(ek, x, w) \rightarrow \pi$  The prove algorithm takes an evaluation key  $ek$  from CRS, and  $(x, w) \in R$  as inputs, and generates a proof  $\pi$ .
- $\text{Verify}(vk, x, \pi) \rightarrow b \in \{0, 1\}$  The verify algorithm takes a verification key  $vk$  from CRS, public input  $x$ , and proof  $\pi$ , and outputs 0(reject) or 1 (accept).
- $\text{Sim}(\text{CRS}, \tau, R)$  The Sim algorithm takes a CRS, a simulation trapdoor  $\tau$ , and a relation  $R$  as input, and returns a proof  $\pi$ .

A zkSNARKs scheme should satisfy the following properties:

- **Completeness:** For any pair  $(x, w) \in R$ , the verifier always accepts the corresponding proof.
- **Knowledge Soundness:** it holds if the prover must know a witness and such knowledge can be efficiently extracted from the prover by using a knowledge extractor.
- **Zero knowledge:** An argument is zero-knowledge if it does not leak any information other than the truth of the statement. There are exist a simulator without secrets can generate valid proofs.
- **Succinctness:** The size of a proof is  $|\pi| \leq \text{poly}(k) \text{polylog}(|x| + |w|)$

**Commit-and-prove SNARK.** Commit-and-prove SNARK is a SNARK(cp-SNARK, for short) [27] that can prove knowledge of  $(x, w)$  such that  $R(x, w) = 1$  holds w.r.t. a witness  $w = (u, w)$  and  $u$  opens a commitment  $c_u$  as follow:

*Definition 3:* We denote a cp-SNARK as a triple of algorithm  $\text{CP} = (\text{KeyGen}, \text{Prove}, \text{VerProof})$ .

- $\text{KeyGen}(ck, R) \rightarrow \text{CRS} = (ek, vk)$  generates the common reference string(CRS).
- $\text{Prove}(ek, x, (c_j)_{j \in [l]}, (u_j)_{j \in [l]}, (o_j)_{j \in [l]}, w) \rightarrow \pi$  outputs the proof of correct commitment.
- $\text{VerProof}(vk, x, (c_j)_{j \in [l]}) \rightarrow b \in \{0, 1\}$  reject or accept the proofs.

The above definition has *perfect completeness*, *computational knowledge soundness* and *zero knowledge* in the random oracle model. Please refer to [27] for more details on the formal definition.

#### IV. PROBLEM STATEMENT

##### A. System Model

As shown in Fig. 1, our framework involves four entities: *server*, *regulators*, *client*, and *bulletin board*. With the involvement of the entities, we consider such a scenario problem: the server in possession of a trained model seeks to convince any later-coming client that the model satisfies a set of fairness metrics typically defined by a group of specialist regulators, while not revealing the model parameters. We note that we consider multiple specialist regulators who hold different fairness metrics dependent on the policy, law regulation, and environment of their domains, so as to ensure an all-around fairness assessment of a trained model. To address the above scenario problem, our framework contains three phases, including the query phase, the auditing phase, and the verification phase, with the following basic workflow:

- The sever commits to the model that will be evaluated, and meanwhile, the regulators commits to their test data that are used for evaluating model fairness.
- Any one regulator can send the test data to the server, and the model is evaluated on the data, thereby obtaining the corresponding evaluation result. Also, the commitment on the evaluation result and the proof regarding evaluation correctness are submitted to the bulletin board. Here refers to the query phase.
- A regulator can audit the model fairness with the evaluation result he obtains, according to the fairness metric he holds. As a result, the regulator submits the auditing result and a proof on correct auditing to the bulletin board. It refers to the auditing phase.
- Any later-coming client who questions the model fairness is able to browse the fairness auditing results and assert the truth with the proofs from the bulletin board. Here refers to the verification phase.

**Remarks.** We remark that the server provides a black-box query inference towards the regulators, without exposing the model parameters.

##### B. Threat Assumptions

We consider either the server or the regulators have the motivation to cheat the client with respect to the model fairness in our scenario problem. We now clarify our concrete threat assumptions on the involved entities.

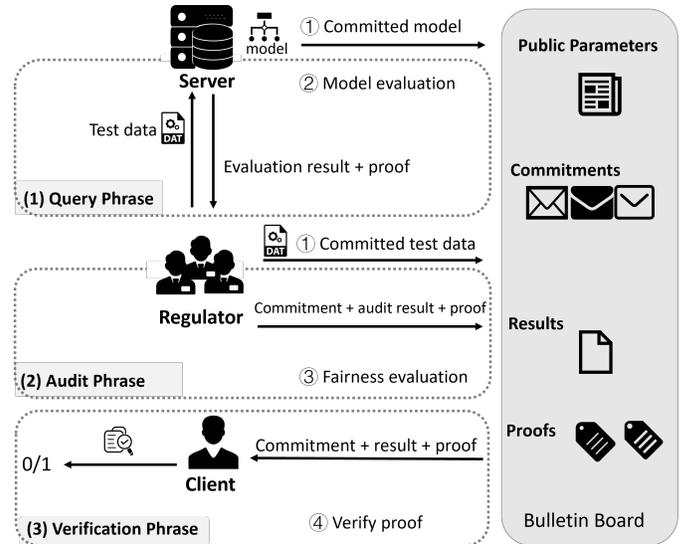


Fig. 1. System overview

**Server:** The server is considered to be *malicious*, which may arbitrarily deviate from the evaluation of model fairness. Specifically, 1) it may give incorrect model predication on test data, such as random values or predication not on the committed model or test data; 2) it also may use an unfair model to interact with the regulator, trying to trick the regulator. We note the server has access to the regulator’s test dataset in plaintext in our scenario. First, given that the source and amount of data is limited, there is a tendency to audit model fairness on publicly available datasets [19] to obtain richer test data. Second, in scenarios where privacy is required for the test dataset, since each of the regulator’s test data only contains sensitive attributes and an identifier as stated in work [14], this represents a lesser privacy risk.

**Regulator:** We consider regulator as *malicious*. The regulator might give wrong fairness audit results, such as auditing an unfair model as fair or vice versa, due to conflict of interest or just machine failure. The regulator also might use test datasets that differs from the previously committed data.

**Client:** We assume client is *honest*. The client can know the algorithms for model prediction and fairness evaluation, but does not have access to the model parameters. The commitments, audit result, and proofs are always available to client.

**Bulletin Board:** We assume that integrity and availability hold for bulletin board. The bulletin board can be instantiated by using a blockchain system, such as Bitcoin or Ethereum. We make the assumption like existing work [45]. We also assume the existence of a secure communication channel between any two entities.

**Remarks.** Note that we do not consider how to train a fair machine learning model and we do not discuss some machine learning attack, such as model extraction, model inversion, and evasion attacks [46] [47]. These studies are outside the scope of this work.

### C. Security Goals

We aim to propose a framework for model fairness auditing in trustless setting, which departs from previous works. The framework establishes an evidence on the server-side model fairness, such that an off-line or later-coming client can be faithfully convinced of the truth respective to model fairness. To be specific, we should achieve the following security goals.

**Trustless Verifiable Model Fairness:** We require that the truth of model fairness can be efficiently verified by any one non-designated client. Concretely, a client of interest as a verifier can verify the public proofs posted by the server and the regulators, so as to determine if the server's model indeed reaches the fairness degree of the regulators over specific test data.

**Model Privacy:** We require that neither evaluation results, auditing results or proofs reveals the private information of the model against any one verifier.

**Audit Integrity:** We can provide the publicly verifiable audit integrity to convince that any later-coming client.

**Accountability:** We require that a verifier can account the misbehavior of the server or the regulator, if evaluation proof or audit proof cannot be verified. This property cannot be provided by MPC-based works.

## V. TECHNICAL CHALLENGE

Above scenario allows the server to know the regulator's test data for evaluating his model, and the regulators to obtain the corresponding evaluation results for auditing model fairness. When any a later-coming client gets the auditing results, we do not desire the client to learn any private information of the server's model, the regulator's test data and the evaluation results. We therefore need to achieve that any client (as non-designated verifier) can check the correctness of model evaluation and fairness auditing without private information of the model, test data and evaluation results.

Zero-knowledge proof technique can be used to address the above scenario problem without leaking private information. As Fig. 1 demonstrated, a regulator can firstly commit to test data (the server also commits to the model). The server then conducts the model evaluation with the test data, yielding the corresponding evaluation results. It also commits to the evaluation results, and generates a proof  $\pi_1$  that the model is indeed evaluated over the test data, with the committed evaluation results as output. After that, the regulator executes the computations of fairness auditing, and generates a proof  $\pi_2$  regarding the execution correctness. Lastly, a verifier checks the validity of both  $\pi_1$  and  $\pi_2$  without knowing the previously mentioned model, test data and evaluation results. Despite the easy-following technical roadmap, we encounter the following two challenges for efficiency:

### A. Supporting Lightweight Verification

Our work adopts the state-of-the-art zkSNARK scheme constructed by Groth [25] (refer to Groth16) due to its short proof size (three group elements) and efficient verification. Despite the optimal performance of the scheme, the direct adoption without any modification cannot satisfy our scenario

requirements. Specifically, to test model fairness and ensure the result reliability require sufficiently large test data, e.g., 6800 test inputs per regulator, as mentioned in [19]. Based on the scenario, the overall proof size will increase linearly with the amount of data, although the single proof of the Groth16 scheme is succinct. But a verifier is considered a thin device with restricted resources, and thus the result proof as described above easily becomes a computational and storage burden on the lightweight verifier. Furthermore, this situation becomes even worse when we want to verify the results of the fairness audit of multiple regulators with different test data and different fairness metrics. Therefore, the challenge is implementing more efficient verification when the model fairness is audited by large test data from multiple regulators.

### B. Improving Proving Performance

Recall that the proof generation process based on an zk-SNARK scheme involves compiling a computation (e.g., model fairness auditing) into a circuit, such as arithmetic circuit and Boolean circuit, and then expressing the circuit with rank-1 constraint system (R1CS) used for generating proof. The involved arithmetic and Boolean operations generally determine the efficiency of the zkSNARK scheme used in practical. For example, the overhead of the Groth16 scheme is friendly for arithmetic operations in the finite field but unfriendly for Boolean operations (due to the radical blow-up in the circuit size required to compile Boolean operations into the arithmetic circuit). Such technical feature, however, is in conflict with the concrete computation of our scenario, since model fairness auditing involves many Boolean-efficient operations such as comparison. Our challenge thus is efficiently handling Boolean operations in proof generation.

## VI. CONCRETE DESIGN

We now present our framework that enables the fairness audit for an ML model while keep model confidentiality in an efficient publicly verifiable manner, enabling each one to assess model fairness individually and thus minimize trust dependency between server, regulator and client. It builds on Groth16 zk-SNARK scheme, Pedersen commitment and signature scheme. As mentioned above, our framework consists of three phase: *query phase*, *audit phase*, and *verification phase*. (1) In the query phase, the regulator query the server's model using the committed test dataset. Then the server evaluate model on test data and generate a proof for correct evaluation, and sends the evaluation results and the evaluation proof to the regulator. (2) The regulator verifies the evaluation proof and evaluate fairness metrics of model. And then the regulator generate auditing proof. And release auditing result and auditing proof to the bulletin board. (3) The client verifies the evaluation proof, audit proof and determines the fairness of the server's model.

### A. Query Phase

1) *Model evaluation.*: In this phase, the main goal of regulator is to obtain the evaluation of the server's model on test data that can be used to audit the model's fairness. Firstly, the regulator collects sufficient test dataset  $\mathcal{X} = \{x_0, \dots, x_{n-1}\}$  (in a legally compliant way or directly extracted from public audit datasets [19]) and commits it to  $\text{cm}_x$  using the Pedersen

commitment scheme. The server also commits its ML model parameter to  $\mathbf{cm}_m$  using the same commitment scheme. Then both parties posting their commitments  $\mathbf{cm}_x$  and  $\mathbf{cm}_m$  to the Bulletin board respectively. Noted that the ML model structure is known to the verifier, we only protect the model parameter privacy, e.g. weight information. Secondly, the regulator transmits test data  $\mathcal{X} = \{x_0, \dots, x_{n-1}\}$  to the server, then the server computes the evaluations of model on the test data  $\hat{\mathcal{Y}} = \{y_0, \dots, y_{n-1}\}$  and generates the evaluation proof  $\pi_e$ . Moreover, the server commit the evaluation of model to  $\mathbf{cm}_y$  and publish it and the evaluation proof  $\pi_e$  to the Bulletin board, and the server sends the model evaluations  $\hat{\mathcal{Y}}$  to the regulator.

2) *Proof generation.*: Our framework conducts the commit-and-prove paradigms [27] so that we can support zero-knowledge evaluation for both secret input and secret models in a straightforward way. Specially, in our scenario, we allow the server get the test data in plaintext, and the regulator obtains the evaluation results for audit the model. However, from the client's perspective, the privacy of the model, test data, and the evaluation result all is preserved.

The claim from [19] stated that an ML model  $M$  is  $\epsilon$ -fair with confidence  $1 - \delta$  if:

$$EFG < \epsilon \wedge \min_{g \in \mathcal{G}} \geq \frac{2}{(\epsilon - EFG)^2} \ln \frac{2|\mathcal{G}||\mathcal{Y}|}{\delta}$$

where  $m_g$  denotes the number of occurrences of  $g$  in  $T$ . Takes  $EFG = 0.05, \epsilon = 0.1, \delta = 0.2$  and  $|\mathcal{G}| = 100$  as example, we need sample number  $m_g \approx 6800$ . In this scenario, although the individual evaluation proof are small, thousands of test data make verifying multiple evaluation proofs expensive. As stated above, however, our goal is to keep the succinctness of proof due to we want to support a lightweight client who has limited memory and computation resources. There are two common techniques to keep succinctness of multiple proofs in literature, one is SNARK recursion [48] [49], and other is proof aggregation [50] [51]. The SNARK recursion can prove the proof is correct, and we can compress a sequence of proofs into one proof. Specially, we can aggregate proofs via recursive composition that create another SNARK for the circuit that contains  $n$  copies of the Groth16 verifier circuit [48]. But the SNARK recursion incur significant practical overhead due to we need to compile the verify algorithm into a circuit, and this is the bottleneck of recursion SNARK efficiency. For example, computing a pairing on the BLS12-377 curve require  $\sim 15000$  constraints [48].

3) *Proof aggregation.*: In our work, we adopt proof aggregation technique. Inspired by SnarkPack [50] [51], we resort to utilize special structure of proof to aggregate multiple proof. SnarkPack propose an approach to reduce the overhead in communication and verification time for verify multiple proofs without the need of further larger trusted setup ceremonies. The SnarkPack allows to aggregate  $n$  Groth16 zk-SNARKs proofs with  $O(\log n)$  proof size and verifier time and can be constructed from two different existing ceremonies (e.g., the "power of tau" for Zcash [38] and Filecoin [52]).

We explain the aggregation protocol used in SnarkPack and how it can adapt to our CP-SNARK scenario below. First, we recall the verification process used in Groth16. A detailed description of Groth16 SNARK protocol can be found in [25].

The proof  $\pi$  in Groth16 consists of three group element  $\pi = (A, B, C) \in \mathbb{G}_1 \times \mathbb{G}_2 \times \mathbb{G}_1$ . For the verification algorithm, we need the verification key vk:

$$vk := \left( g^\alpha, h^\beta, \left\{ S_j = g^{\frac{\beta_j(s) + \alpha w_j(s) + y_j(s)}{\gamma}} \right\}_{j=0}^t, h^\gamma, h^\delta \right)$$

The verifier need check that pairing equations is satisfy:  $e(A, B) = e(g^\alpha, h^\beta) \cdot e(\sum_{j=0}^t S_j^{\alpha_j}, h^\gamma) \cdot e(C, h^\delta)$ , where  $[a]$  is public input.

The high level idea of Groth16 aggregation is straightforward: instead of checking that  $n$  different pairing equations are simultaneously satisfied, it is sufficient to prove that only one inner pairing product of a random linear combination of these equations defined by a verifier's random challenge  $r \in \mathbb{Z}_p$  holds. The same idea is heavily exploited in polynomial commitment, SNARK batch, SNARK recursive [48]. Specially, consider  $n$  proof  $[\pi_i]_{i=0}^{n-1} = \{A_i, B_i, C_i\}_{i=0}^{n-1}$ , the verifier need to check  $n$  equations of  $e(A_i, B_i) = Y_i \cdot e(C_i, h^\delta)$ , where

$$Y_i = e(g^\alpha, h^\beta)^n \cdot e\left(\prod_{i=0}^{n-1} S_i^{\sum_{j=0}^{n-1} a_{i,j} \cdot r^j}, h^\gamma\right)$$

The aggregation will instead check a single randomized equation:

$$\prod_{i=0}^{n-1} e(A_i, B_i)^{r^i} = \prod_{i=0}^{n-1} Y_i^{r^i} \cdot e\left(\prod_{i=0}^{n-1} C_i^{r^i}, h^\delta\right)$$

And we rewritten above equation as:

$$Z_{AB} = Y'_{prod} \cdot e(Z_C, h^\delta)$$

where  $Z_{AB} = \prod_{i=0}^{n-1} e(A_i, B_i)^{r^i}$ ,  $Z_C = \prod_{i=0}^{n-1} C_i^{r^i}$  and  $Y'_{prod} = \prod_{i=0}^{n-1} Y_i^{r^i}$ . And then we will check that  $Z_{AB}, Z_C$  are consistent with the initial proof triples. Here we use two notions: the target inner pairing product (TIPP) and the multi-exponentiation inner product (MIPP) (detail can see [50]).

- TIPP: takes some committed vector  $\mathbf{A} \in \mathbb{G}_1^n, \mathbf{B} \in \mathbb{G}_2^n$  and shows that  $Z_{AB} = \prod_{i=0}^{n-1} e(A_i, B_i)$ ;
- MIPP: takes a committed vector  $\mathbf{C} \in \mathbb{G}_1^n$  and a vector  $\mathbf{r} \in \mathbb{Z}_p^n$  and shows that  $Z_C = \prod_{i=0}^{n-1} C_i^{r^i}$

After that we can use TIPP and MIPP to generate the aggregated proof  $\pi = (\pi_t, \pi_m, T_{AB}, U_{AB}, T_C, U_C, Z_{AB}, Z_C)$ , where the last two elements are required to verify the Groth16 equation, the first two elements used to verify the TIPP and MIPP arguments, and other elements are required for the verifier derive randomness  $r$  in Fiat-Shamir transformation [53]. After verify TIPP and MIPP proof, the regulator use  $Z_{AB}, Z_C$  as the linear combination of the proofs. Then the regulator verify the Groth16 equation using the aggregated proof  $Z_{AB}, Z_C$  and decides whether to move to the next stage.

In the case of CP-SNARK, we need additional element  $D$  of the proof that contains a commitment to the data and to create a CPlink to link  $D$  to the external commitment. Also we need some additional element in CRS to create  $D$  and the CPlink. Nevertheless, the special structure of the proof does

TABLE I. CONFUSION MATRIX

	Actual-Positive	Actual-Negative
Evaluation-Positive	True Positive (TP) $TPR = \frac{TP}{TP+FN}$ $PPV = \frac{TP}{TP+FP}$	False Positive (FP) $FPR = \frac{FP}{FP+TN}$
Evaluation-Negative	False Negative (FN) $FNR = \frac{FN}{TP+FN}$	True Negative (TN) $TNR = \frac{TN}{TN+FP}$

not change. In order to verify the proof, we only need verify equation of the structure:  $e(A, B) = Y \cdot e(C, h^\delta)$ . Thus, we can obtain the aggregated proof in the same way.

### B. Auditing Phase

In this phase, multiple regulators aim to evaluate particular fairness metrics depending on the environment, policy, and industry to determine the fairness of the ML model.

1) *Fairness audit*: Following the common approach in [42], we can use a confusion matrix (see Table I) to compute statistical metrics of ML model, which is what most statistical measures of fairness rely on. The basic notion in confusion matrix as follow:

- True positive (TN): a case when the predicted and actual evaluation are both in the positive class.
- False positive (FP): a case predicted to be in the positive class when the actual outcome belongs to the negative class.
- False negative (FN): a case predicted to be in the negative class when the actual outcome belongs to the positive class.
- True negative (TN): a case when the predicted and actual evaluation are both in the negative class.

Based on these basic concepts, we calculate the fairness metrics. Specifically, the fairness notion we took in this work is as follows: Demographic parity (DP), Equalize odds (EO), Equal opportunity, and Disparate impact (DI). Use these formulation, following the basic paradigm in zk-SNARKs, we can express the whole fairness computation as arithmetic circuit and then generate proof.

2) *Technical observation*: As mentioned earlier, the challenge in generating proof is that we must handle many non-linear operations because we adopt group-based fairness notions, e.g., divides, comparison, sorting, and so on. To resolve the challenge, our framework uses the observation that for a prover to convince a verifier that it knows the output of some non-linear operation, the prover does not actually need to execute the non-linear operation in the circuit. Instead, the prover just needs to prove that the output of the non-linear operations is correct. For example, suppose the prover wants to prove  $z = x/y$  to the verifier. In that case, the prover does not need to straight to compiler divide to a circuit but simply provides a divided result  $z$  as an "advice" and then prove multiplication operation  $x = z \times y$ . Note that the multiplication operation only contain one multiplication gate, so it much more efficient than naive encoding the divide operation to

a circuit (e.g., compute inverse use Fermat's little theorem,  $a \cdot a^{p-2} = 1$ ). The observation is broadly used in literature [24], [39], [54], [55]. Next, we will show how to bring this idea into our framework to help model audit.

3) *Handle non-linear operations*: At a high level, we can split the computation in the audit phase into two-component: one is to evaluate fairness metrics, and the other is to compute the empirical fairness gap. First, we summarize the typical non-linear operation in the auditing process below and then design a protocol to prompt efficiency. The typical operations in statistical fairness evaluate as follow: 1) divides; 2) absolute value; 3) comparison; and 4) maximum or minimum value. To clarify the research methodology, we consider the following examples. First, the regulator needs to compute the basic metrics: TP, FP, TN, and FN. And then calculate fairness metrics like TPR and FRP. Finally, the regulator computes the EFG corresponding fairness notion, such as DP.

For division operations, we have solved as above. For comparison operation, e.g.  $x \geq 0$ , we ask the prover to provide the bit decomposition  $(a_0, \dots, a_k)$  of  $x$  as a witness ( $a_k$  denote sign bit, e.g. positive(1) or negative(0)). Then we can check that: 1) each  $a_i$  is binary:  $a_i(a_i - 1) = 0, \forall i \in [k]$ ; 2) The correctness of bit-decomposition of  $x$ :  $a_k(x - \sum_{i=0}^{k-1} a_i 2^i) + (1 - a_k)(x + \sum_{i=0}^{k-1} a_i 2^i) = 0$ .

In calculating the EFG, we need to find the maximum metrics gap between different group. Naive computation is not feasible because it requires us to make a two-by-two difference between all the elements in the group and then select the maximum gap value. A more thoughtful way is to sort the array and then use the maximum value minus the minimum value to get EFG. Nevertheless, directly representing a sorting algorithm such as QuickSort as a circuit requires a comparison that contains  $O(n \log n)$ , which is expensive. Following the above observation, we asked the prover to provide some "advice" as a witness to improve efficiency, and we can combine absolute and maximum value into one relation. Concretely, assume that we have  $(x_0, \dots, x_{n-1})$  denote the fairness metrics on  $n$  test data and want to compute  $\max |x_i - x_j|$  between different group. The prover is required to provide the maximum  $x_{max}$  and the minimum value  $x_{min}$  in list as an auxiliary witness. Then we can check that  $x_{max}$  is actually the maximum number as follows:

- 1)  $x_{max} - x_j \geq 0 \forall j \in [n]$ .
- 2)  $\exists j \in [n]$  such that  $x_{max} - x_j = 0$ . This condition is equivalent to  $x_{max} \cdot \prod_{j=0}^{n-1} (x_{max} - x_j) = 0$ .

The first condition can be checked by bit-decomposing of  $x_{max} - x_j$  and then checks are exactly the same as the comparison operation. Similarly, the check of  $x_{min}$  is exactly the same as  $x_{max}$  above.

Overall, comparing to straight compute maximize value above, the prover only additionally provides bit-decomposing of maximize and minimize value, and the protocol checks two additional bit decomposition. After all of computation done, the regulator publish all of the commitments, proofs, and audit result with associated signature on bullet board.

TABLE II. EVALUATION PROOF TIME OF FLR MODEL

Dataset	preprocess time(s)	prover time(s)	verifier time(s)
German	2.139	0.736	0.0159
Bank	110.853	34.9684	0.0293
Adult	82.272	26.921	0.0403

### C. Verification Phase

Depending on the application area, regulator is selected to verify the correctness of the audit results. After get the audit result from  $y$ , the client request committed values for the ML models  $cm_m$ , test data  $cm_x$ , and evaluation results  $cm_y$ . The client then request evaluation proof  $\pi_e$  and auditing proof  $\pi_a$  from the bulletin board.

Upon reception of all of message, the client verifies the signature using the public key of the selected regulators and server for the authenticity of commitment, proof and results. If satisfied, client then verifies the evaluation proof  $\pi_e$  and audit proof  $\pi_a$  using the verification key and commitment. If all these verification pass, then the client will be convinced by the regulator's auditing results and thus determinate the fairness of the model.

If the client verify failure then one can identify the misbehaving party and take penalize it, in the form of reputation evaluation, incentives, etc. Then the party responsible for the misleading behavior is deterred and restart a new auditing process. After all checks are successful, the client can determine the correctness of the ML model auditing results.

## VII. EVALUATION

We implemented our fairness audit framework and we present the experiment result in this section.

### A. Setup

We have implemented the fairness audit framework in C++ using the libsnark [56] library. We run all of the experiments on 4-core Intel i7-5600k (2.6 GHZ, 8 physical cores) and 48 GB of RAM with Ubuntu18.04. Note that we run our experiments on Docker container. Our current implementation is only use a single CPU core. We report the time in seconds and take the average of 10 runs per experiment as the result.

We used three real-world datasets from various domains: German credit dataset (German), Bank marketing dataset (Bank), and Adult income dataset (Adult) [57]. The datasets vary in size and disparity of minority groups and as such some can be used to create fair or unfair models. In this experiment, we train fair logistic regression (FLR)<sup>1</sup>. Note that due to zk-SNARK systems only support group elements, we use a generic 8-bit unsigned quantization technique to transform float into integers.

<sup>1</sup>The implementation of fair machine learning model is based on the repository <https://github.com/mbilalzafar/fair-classification>.

### B. Performance of Model Query

In this section, we report the performance of the query phase of the model. First, we use FLR model to measure the time to generate evaluation proofs for all three datasets. Since matrix multiplication is involved in all three models, a representative FLR is selected to measure the performance. Table II shows the results of evaluation proof in model query phase. The FLR model needs matrix-vector multiplication between the weight vector and input matrix of test data, where the sizes of input matrix are  $1000 \times 20$  and  $45550 \times 20$ , and  $48880 \times 15$ . As a result, the input test data and the number of samples affect the circuit constraints and thus the performance.

After that we measure the overhead incurred by aggregating the proofs. Due to the different sizes of the three datasets to ensure the fairness of the model, we set different fairness gaps and confidence levels just like in [19]. We performed our test on the mentioned datasets with  $\delta = 0.05$  and  $\epsilon = 0.1$ . Our evaluation result can see in Fig. 2. The results show that aggregated proofs of thousands of proofs are relatively small in size and can be verified in a few seconds. Although the proofs take more time, this is not impractical. In practice, aggregation operations can be performed offline without affecting the online use of the model after deployment.

### C. Performance of Fairness Audit

After query phase, the regulator calculates fairness metrics based on the evaluation results from the server. Thus, we measure computation times of fairness metrics in three datasets. Naturally, we split the data into a training and test subset. The sizes of the three datasets are 200, 9100 and 9000, respectively. In Bank dataset, we use the marital status feature as binary sensitive attribute and income for labels. We considered gender as a binary sensitive variable in Adult and German. Table III shows the evaluation results of fairness audit. The experimental results prove the usefulness of our framework, which takes only a few milliseconds to verify the correctness of the fairness audit results, making it easy for anyone to ensure that they are protected from discrimination. In addition, both preprocessing and prove times are within acceptable bounds.

## VIII. DISCUSSION AND FUTURE WORK

One problem with the Groth16 system used in the framework instantiation, despite its state-of-the-art proof scale, is that it requires a trusted third party to generate the CRS used to construct the proofs. While it is possible to use MPC to generate reliable CRS [58], a transparent zkp system [28] could be used instead to avoid trusted party, which is left for future exploration. We expect that better ZK systems will emerge to replace the ZKP schemes we use and thus improve the efficiency of proposed framework. In order to achieve optimal ZK systems in all aspects, such as proof size, proofing time, etc., one promising direction is proof composition [37], [59]. Also, we can naturally extend our work to support confidential model prediction and model accuracy assessment [39], [40].

For future work, we want to explore further protecting the privacy of test data on the server. Especially since there are no publicly available test data and the user data is susceptible in some scenarios. A promising direction is the use of verifiable

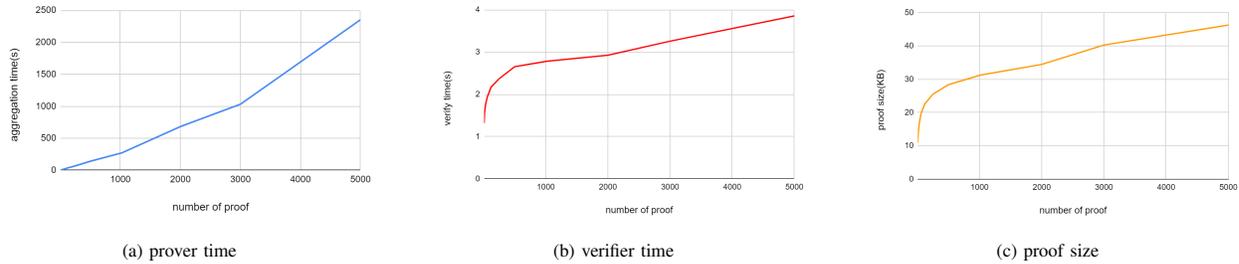


Fig. 2. Performance of Aggregation

TABLE III. FAIRNESS AUDIT PERFORMANCE OF FLR MODEL ON THREE DATASETS

Dataset	preprocess time(s)				prover time(s)				verifier time(s)			
	DP	EO	EOP	DI	DP	EO	EOP	DI	DP	EO	EOP	DI
German	8.17	8.62	9.84	8.29	1.64	2.35	1.95	2.25	0.0062	0.0059	0.0061	0.0063
Bank	87.13	86.42	86.01	88.23	25.55	23.23	24.66	25.28	0.0052	0.0061	0.0063	0.0052
Adult	81.19	80.58	81.56	82.48	23.23	24.96	23.86	23.75	0.0067	0.0068	0.0069	0.0065

encryption techniques [60] to ensure integrity and confidentiality. Finally, it is also interesting to explore other fairness definitions beyond group-based as a research direction.

## IX. CONCLUSION

This work proposes a framework to prompt publicly fair audits for a machine learning model. Unlike previous work, our construction only assumes the third-party collection test data and does not rely on the third party. We minimize the trust between the server, the regulator, and the client. Also, our framework can support multiple regulators to provide more strength and border fairness without additional trust assumptions. Our experimental evaluation confirms that our framework is practical for fairness auditing ML models with real datasets.

## ACKNOWLEDGMENT

The authors would like to thank Dr. Jiasi Weng for valuable and helpful discussions on this work. Wuzheng Tan was supported by National Natural Science Foundation of China (Grant Nos. 62272199).

## REFERENCES

- [1] S. M. D. A. C. Jayatilake and G. U. Ganegoda, "Involvement of Machine Learning Tools in Healthcare Decision Making," *Journal of Healthcare Engineering*, vol. 2021, p. e6679512, 2021.
- [2] T. Tulabandhula and C. Rudin, "On combining machine learning with decision making," *Machine Learning*, vol. 97, no. 1, pp. 33–64, 2014.
- [3] A. E. W. Johnson, M. M. Ghassemi, S. Nemati, K. E. Niehaus, D. A. Clifton, and G. D. Clifford, "Machine Learning and Decision Support in Critical Care," *Proceedings of the IEEE*, vol. 104, no. 2, pp. 444–466, 2016.
- [4] T. Kamishima, S. Akaho, H. Asoh, and J. Sakuma, "Fairness-Aware Classifier with Prejudice Remover Regularizer," in *Machine Learning and Knowledge Discovery in Databases*, ser. Lecture Notes in Computer Science. Springer, 2012, pp. 35–50.
- [5] M. Hardt, E. Price, E. Price, and N. Srebro, "Equality of Opportunity in Supervised Learning," in *Advances in Neural Information Processing Systems*, vol. 29. Curran Associates, Inc., 2016.
- [6] G. Pleiss, M. Raghavan, F. Wu, J. Kleinberg, and K. Q. Weinberger, "On Fairness and Calibration," in *Advances in Neural Information Processing Systems*, vol. 30. Curran Associates, Inc., 2017.
- [7] J.-G. Lee, Y. Roh, H. Song, and S. E. Whang, "Machine Learning Robustness, Fairness, and their Convergence," in *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, ser. KDD '21. Association for Computing Machinery, 2021, pp. 4046–4047.
- [8] H. Zhang, X. Chu, A. Asudeh, and S. B. Navathe, "OmniFair: A Declarative System for Model-Agnostic Group Fairness in Machine Learning," in *Proceedings of the 2021 International Conference on Management of Data*, ser. SIGMOD '21. Association for Computing Machinery, 2021, pp. 2076–2088.
- [9] R. K. E. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilović, S. Nagar, K. N. Ramamurthy, J. Richards, D. Saha, P. Sattigeri, M. Singh, K. R. Varshney, and Y. Zhang, "AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias," *IBM Journal of Research and Development*, vol. 63, no. 4/5, pp. 4:1–4:15, 2019.
- [10] B. d'Alessandro, C. O'Neil, and T. LaGatta, "Conscientious Classification: A Data Scientist's Guide to Discrimination-Aware Classification," *Big Data*, vol. 5, no. 2, pp. 120–134, 2017.
- [11] B. H. Zhang, B. Lemoine, and M. Mitchell, "Mitigating Unwanted Biases with Adversarial Learning," in *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, ser. AIES '18. Association for Computing Machinery, 2018, pp. 335–340.
- [12] R. Berk, H. Heidari, S. Jabbari, M. Joseph, M. Kearns, J. Morgenstern, S. Neel, and A. Roth, "A Convex Framework for Fair Regression," 2017.
- [13] T. Bolukbasi, K.-W. Chang, J. Y. Zou, V. Saligrama, and A. T. Kalai, "Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings," in *Advances in Neural Information Processing Systems*, vol. 29. Curran Associates, Inc., 2016.
- [14] M. Veale and R. Binns, "Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data," *Big Data & Society*, vol. 4, no. 2, p. 2053951717743530, 2017.
- [15] N. L. Martinez, M. A. Bertran, A. Papadaki, M. Rodrigues, and G. Sapiro, "Blind Pareto Fairness and Subgroup Robustness," in *Proceedings of the 38th International Conference on Machine Learning*. PMLR, 2021, pp. 7492–7501.
- [16] P. Lahoti, A. Beutel, J. Chen, K. Lee, F. Prost, N. Thain, X. Wang, and E. Chi, "Fairness without Demographics through Adversarially Reweighted Learning," in *Advances in Neural Information Processing Systems*, vol. 33. Curran Associates, Inc., 2020, pp. 728–740.

- [17] S. Park, S. Kim, and Y.-s. Lim, "Fairness Audit of Machine Learning Models with Confidential Computing," in *Proceedings of the ACM Web Conference 2022*, ser. WWW '22. Association for Computing Machinery, 2022, pp. 3488–3499.
- [18] N. Kilbertus, A. Gascon, M. Kusner, M. Veale, K. Gummadi, and A. Weller, "Blind Justice: Fairness with Encrypted Sensitive Attributes," in *Proceedings of the 35th International Conference on Machine Learning*. PMLR, 2018, pp. 2630–2639.
- [19] S. Segal, Y. Adi, B. Pinkas, C. Baum, C. Ganesh, and J. Keshet, "Fairness in the Eyes of the Data: Certifying Machine-Learning Models," in *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, ser. AIES '21. Association for Computing Machinery, 2021, pp. 926–935.
- [20] S. Pentyala, D. Melanson, M. De Cock, and G. Farnadi, "PrivFair: A Library for Privacy-Preserving Fairness Auditing," 2022.
- [21] T. Cloosters, M. Rodler, and L. Davi, "TEEREX: Discovery and Exploitation of Memory Corruption Vulnerabilities in SGX Enclaves," in *29th USENIX Security Symposium*, 2020.
- [22] T. Cloosters, J. Willbold, T. Holz, and L. Davi, "SGXFuzz: Efficiently Synthesizing Nested Structures for SGX Enclave Fuzzing," in *USENIX Security*, 2022.
- [23] N. Bitansky, R. Canetti, A. Chiesa, and E. Tromer, "From extractable collision resistance to succinct non-interactive arguments of knowledge, and back again," in *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, ser. ITCS '12. Association for Computing Machinery, 2012, pp. 326–349.
- [24] B. Parno, J. Howell, C. Gentry, and M. Raykova, "Pinocchio: Nearly Practical Verifiable Computation," in *2013 IEEE Symposium on Security and Privacy*, 2013, pp. 238–252.
- [25] J. Groth, "On the Size of Pairing-Based Non-interactive Arguments," in *Advances in Cryptology – EUROCRYPT 2016*, ser. Lecture Notes in Computer Science. Springer, 2016, pp. 305–326.
- [26] A. Gabizon, Z. J. Williamson, and O. Ciobotaru, "PLONK: Permutations over Lagrange-bases for Oecumenical Noninteractive arguments of Knowledge," 2019.
- [27] M. Campanelli, D. Fiore, and A. Querol, "LegoSNARK: Modular Design and Composition of Succinct Zero-Knowledge Proofs," in *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, ser. CCS '19. Association for Computing Machinery, 2019, pp. 2075–2092.
- [28] S. Setty, "Spartan: Efficient and General-Purpose zkSNARKs Without Trusted Setup," in *Advances in Cryptology – CRYPTO 2020*, ser. Lecture Notes in Computer Science. Springer International Publishing, 2020, pp. 704–737.
- [29] J. Bootle, A. Chiesa, Y. Hu, and M. Orrú, "Gemini: Elastic SNARKs for Diverse Environments," in *Advances in Cryptology – EUROCRYPT 2022*, ser. Lecture Notes in Computer Science. Springer International Publishing, 2022, pp. 427–457.
- [30] S. Goldwasser, S. Micali, and C. Rackoff, "The knowledge complexity of interactive proof-systems," in *Proceedings of the Seventeenth Annual ACM Symposium on Theory of Computing - STOC '85*. ACM Press, 1985, pp. 291–304.
- [31] J. Kilian, "A note on efficient zero-knowledge proofs and arguments (extended abstract)," in *Proceedings of the Twenty-Fourth Annual ACM Symposium on Theory of Computing*, ser. STOC '92. Association for Computing Machinery, 1992, pp. 723–732.
- [32] S. Micali, "CS proofs," in *Proceedings 35th Annual Symposium on Foundations of Computer Science*, 1994, pp. 436–453.
- [33] B. Bünz, J. Bootle, D. Boneh, A. Poelstra, P. Wuille, and G. Maxwell, "Bulletproofs: Short Proofs for Confidential Transactions and More," in *2018 IEEE Symposium on Security and Privacy (SP)*, 2018, pp. 315–334.
- [34] T. Xie, J. Zhang, Y. Zhang, C. Papamanthou, and D. Song, "Libra: Succinct Zero-Knowledge Proofs with Optimal Prover Computation," in *Advances in Cryptology – CRYPTO 2019*, ser. Lecture Notes in Computer Science. Springer International Publishing, 2019, pp. 733–764.
- [35] J. Zhang, T. Liu, W. Wang, Y. Zhang, D. Song, X. Xie, and Y. Zhang, "Doubly Efficient Interactive Proofs for General Arithmetic Circuits with Linear Prover Time," in *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, ser. CCS '21. Association for Computing Machinery, 2021, pp. 159–177.
- [36] E. Ben-Sasson, I. Bentov, Y. Horesh, and M. Riabzev, "Scalable Zero Knowledge with No Trusted Setup," in *Advances in Cryptology – CRYPTO 2019*, ser. Lecture Notes in Computer Science. Springer International Publishing, 2019, pp. 701–732.
- [37] A. Golovnev, J. Lee, S. Setty, J. Thaler, and R. S. Wahby, "Brakedown: Linear-time and post-quantum SNARKs for R1CS," 2021.
- [38] E. Ben Sasson, A. Chiesa, C. Garman, M. Green, I. Miers, E. Tromer, and M. Virza, "Zerocash: Decentralized Anonymous Payments from Bitcoin," in *2014 IEEE Symposium on Security and Privacy*, 2014, pp. 459–474.
- [39] T. Liu, X. Xie, and Y. Zhang, "zkCNN: Zero Knowledge Proofs for Convolutional Neural Network Predictions and Accuracy," in *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, ser. CCS '21. Association for Computing Machinery, 2021, pp. 2968–2985.
- [40] J. Weng, J. Weng, G. Tang, A. Yang, M. Li, and J.-N. Liu, "pvCNN: Privacy-Preserving and Verifiable Convolutional Neural Network Testing," 2022.
- [41] P. Grubbs, A. Arun, Y. Zhang, J. Bonneau, and M. Walfish, "{Zero-Knowledge} Middleboxes," in *31st USENIX Security Symposium (USENIX Security 22)*, 2022, pp. 4255–4272.
- [42] S. Verma and J. Rubin, "Fairness Definitions Explained," in *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)*, 2018, pp. 1–7.
- [43] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, "Fairness through awareness," in *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, ser. ITCS '12. Association for Computing Machinery, 2012, pp. 214–226.
- [44] A. Chouldechova, "Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments," *Big Data*, vol. 5, no. 2, pp. 153–163, 2017.
- [45] S. Kanjalkar, Y. Zhang, S. Gandlur, and A. Miller, "Publicly Auditible MPC-as-a-Service with succinct verification and universal setup," in *2021 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*, 2021, pp. 386–411.
- [46] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership Inference Attacks Against Machine Learning Models," in *2017 IEEE Symposium on Security and Privacy (SP)*, 2017, pp. 3–18.
- [47] Y. Zhang, R. Jia, H. Pei, W. Wang, B. Li, and D. Song, "The Secret Revealer: Generative Model-Inversion Attacks Against Deep Neural Networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 253–261.
- [48] S. Bowe, A. Chiesa, M. Green, I. Miers, P. Mishra, and H. Wu, "ZEXE: Enabling Decentralized Private Computation," in *2020 IEEE Symposium on Security and Privacy (SP)*, 2020, pp. 947–964.
- [49] N. Tyagi, B. Fisch, A. Zitek, J. Bonneau, and S. Tessaro, "VerRSA: Verifiable Registries with Efficient Client Audits from RSA Authenticated Dictionaries," in *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*. ACM, 2022, pp. 2793–2807.
- [50] B. Bünz, M. Maller, P. Mishra, N. Tyagi, and P. Vesely, "Proofs for Inner Pairing Products and Applications," in *Advances in Cryptology – ASIACRYPT 2021*, ser. Lecture Notes in Computer Science. Springer International Publishing, 2021, pp. 65–97.
- [51] N. Gailly, M. Maller, and A. Nitulescu, "SnarkPack: Practical SNARK Aggregation," in *Financial Cryptography and Data Security*, ser. Lecture Notes in Computer Science. Springer International Publishing, 2022, pp. 203–229.
- [52] "Filecoin: A decentralized market for storage," <https://filecoin.io>.
- [53] A. Fiat and A. Shamir, "How To Prove Yourself: Practical Solutions to Identification and Signature Problems," in *Advances in Cryptology – CRYPTO '86*, ser. Lecture Notes in Computer Science. Springer, 1987, pp. 186–194.
- [54] Y. Zhang, D. Genkin, J. Katz, D. Papadopoulos, and C. Papamanthou, "vSQL: Verifying Arbitrary SQL Queries over Dynamic Outsourced Databases," in *2017 IEEE Symposium on Security and Privacy (SP)*, 2017, pp. 863–880.

- [55] S. Angel, A. J. Blumberg, E. Ioannidis, and J. Woods, "Efficient Representation of Numerical Optimization Problems for {SNARKs}," in *31st USENIX Security Symposium (USENIX Security 22)*, 2022, pp. 4273–4290.
- [56] "libsnark: a c++ library for zksnark proofs," <https://github.com/sciprnlab/libsnark>.
- [57] D. Dua and C. Graff, "UCI machine learning repository," 2017. [Online]. Available: <http://archive.ics.uci.edu/ml>
- [58] E. Ben-Sasson, A. Chiesa, M. Green, E. Tromer, and M. Virza, "Secure Sampling of Public Parameters for Succinct Zero Knowledge Proofs," in *2015 IEEE Symposium on Security and Privacy*, pp. 287–304.
- [59] T. Xie, Y. Zhang, and D. Song, "Orion: Zero Knowledge Proof with Linear Prover Time," in *Advances in Cryptology – CRYPTO 2022*, ser. Lecture Notes in Computer Science. Springer Nature Switzerland, 2022, pp. 299–328.
- [60] D. Fiore, A. Nitulescu, and D. Pointcheval, "Boosting Verifiable Computation on Encrypted Data," in *Public-Key Cryptography – PKC 2020*, ser. Lecture Notes in Computer Science. Springer International Publishing, pp. 124–154.

# An OCR Engine for Printed Receipt Images using Deep Learning Techniques

Çağrı Sayallar<sup>1</sup>, Ahmet Sayar<sup>2</sup>, Nurcan Babalık<sup>3</sup>  
Nacsoft, Kocaeli University Technopark<sup>1,3</sup>  
Computer Engineering, Kocaeli University<sup>2</sup>  
Kocaeli, Turkey

**Abstract**—The digitization of receipts and invoices, and the recording of expenses in industry and accounting have begun to be used in the field of finance tracking. However, 100% success in character recognition for document digitization has not yet been achieved. In this study, a new Optical Character Recognition (OCR) engine called Nacsoft OCR was developed on Turkish receipt data by using artificial intelligence methods. The proposed OCR engine has been compared to widely used engines, Easy OCR, Tesseract OCR, and the Google Vision API. The benchmarking was made on English and Turkish receipts, and the accuracies of OCR engines in terms of character recognition and their speeds are presented. It is known that OCR character recognition engines perform better at word recognition when provided word position information. Therefore, the performance of the Nacsoft OCR engine in determining the word position was also compared with the performance of the other OCR engines, and the results were presented.

**Keywords**—Optical Character Recognition (OCR); image processing; deep learning; benchmarking; receipt

## I. INTRODUCTION

With the introduction of computers into our lives, important documents for the user began to be stored in the computer environment. Although most documents are stored electronically, there are still printed-paper documents that we frequently use in daily life. Invoices and receipts which are among such documents printed on nondurable paper contains information that needs to be saved. When it comes to saving and storing a paper document, the first thing that comes to mind is to scan the document and store the document image in electronic environment. With this method, documents can be stored properly by gaining space. Nevertheless, operations such as listing, sorting and processing these document images are carried out by people, which means a loss of time and resources, especially for companies. In order to meet this need, Optical Character Recognition (OCR) engines have been developed to automate the processing of document images. OCR engines have been developed to read the images containing the text and convert them into processable text outputs.

The most obvious examples of digitized printed documents are receipts and invoices. Receipts and invoices carry data on them that may be important to the user, such as amount, tax, date. The user may wish to store or process this information. Applications have been developed to meet these needs of the user. The common purpose of these applications is to enable the user to track his or her or someone else's spending, store and rank their spending. These applications may encounter many problems when reading the receipt data. Since receipts

and invoices are made of paper, they are nondurable, so they wrinkle very quickly, wear out in a short time and the writing on them can be easily erased. In addition, factors such as different fonts, images, shapes, presentation of information in the form of tables, background of the receipt image, and oblique withdrawal of the receipt make the situation more difficult. These factors are examples of problems that make it difficult to read a receipt. In the applications mentioned, it is necessary to use the OCR engine, which gives the best result despite these problems. The field of OCR is a field that attracts a lot of attention and is competitive. For this reason, there are many OCR engines produced. Since OCR engines are mostly useful for large companies, most OCR engines produced are for commercial purposes and the methods used are not shared. The most well-known of the commercially produced OCR engines are Amazon Textract and Google Vision Api. Unlike commercially produced OCR engines, there are also a small number of open source OCR engines. Examples of open source OCR engines are Tesseract [1], Easy OCR [2] and OCRopus [3].

The aim of this study is to develop an open source OCR engine using artificial intelligence methods. The developed OCR engine is trained using receipt data. The receipt dataset is difficult to find because the receipt images contain personal information belonging to the user. For this reason, we need to create the dataset to be used for training ourselves. Since all of the data obtained by us consisted of Turkish receipts, Turkish receipt images were used for the training. The success of the developed OCR engine is compared to other existing OCR engines. For comparison, both Turkish receipt dataset and English receipt dataset were used. Since the training dataset consists of Turkish receipt images, the comparison of the achievements of OCR engines was also made on Turkish receipts. As mentioned earlier, these receipts contain the personal information of the owner on them. For this reason, the Turkish dataset and the results of the comparison cannot be shared. Both for the sharing of the comparison results and to measure the success of the developed OCR engine in different datasets, the comparison is also made on the English receipt data. Test images from the ICDAR-2019 SROIE [4] competition were used for the English receipt dataset.

Although this study focuses on only one document type, the methods used can be applied to other document types. The method used in the study consists of three main headings. These; pre-preparation, word detection and word reading. In the preliminary preparation stage, the receipt areas are determined in the visual by passing through the visual masking and

segmentation model that comes first. Then, the corner detection algorithm is operated on the obtained mask and the perspective process is applied by detecting the four corners of the receipt. Thus, the receipt image is freed from unnecessary background and more focus is provided on the regions containing text. In the next step, the object detection algorithm is applied to the image obtained and the locations of the words in the receipt are determined. In the final stage, the words detected are passed through the Convolutional Recurrent Neural Network (C-RNN) [5] model and the reading process is performed. In this way, the incoming receipt image is OCR'd and the text output is produced. The success of the OCR engine developed in this study is compared to Tesseract, Easy OCR and Google Vision Api. When making the comparison, the test dataset consisting of Turkish receipt images selected by us and not used during the training phase and also English receipt data from the ICDAR2019-SROIE competition is used. Comparisons are evaluated on the basis of speed and accuracy, and the results from each dataset are shown in separate tables.

The flow in this article can be summarized as follows. In the following part of the study, the OCR engine developed will be explained. The method and dataset used in this regard will be detailed. In Section III, the success of the proposed OCR engine will be detailed from the dataset used to compare it with other commonly used OCR engines and the benchmark metrics used. Section IV contains the results from the comparison, interpretations and analyses of these results. Section V briefly summarizes our work and talks about the work to be done in the future.

## II. RELATED WORK

OCR engines have been a topic of interest for a long time. Even though studies have been carried out for a long time, the OCR problem is still not a solved problem. However, there are OCR engines that can achieve high success compared to others. Applications such as Amazon Textract and Google Vision Api, which are among these OCR engines, are chargeable to use. Since OCR engines are generally produced for commercial purposes, very few open source OCR engines are available. The most well-known open source OCR engine in the literature is Tesseract. Tesseract OCR engine differs from our study in terms of the methods it uses. While deep learning methods are used for OCR in our study, Tesseract performs OCR with pixel operations. The Tesseract OCR engine first determines the text areas in the image by performing page layout analysis. Blobs are obtained by applying connected component analysis in the specified text fields. The detected blobs are then separated into lines and words. After the words are divided into characters with two different methods, the text recognition process is performed using the two-pass adaptive classifier.

Apart from the Tesseract OCR engine, there is another open source OCR engine called OCRopus. This OCR engine, first determines the text fields by page layout analysis like Tesseract. Then the text fields determined by the page layout are sent to the Text Line Recognition stage, and the language of the text and the writing direction of the text (right to left, left to right) are determined. It uses dynamic programming algorithm for character detection and multi-layer perceptrons (MLPs) for character recognition. There is also a study [6], which uses slightly more modern methods and whose text

recognition stage is similar to our work. In the text detection phase, all contours are detected with the Canny algorithm and then lines are determined using these contours. Then, preprocessing operations such as Noise removal, perspective correction, baseline correction are applied to improve the reading process before the text recognition stage. The lines detected in the text detection stage are divided into smaller pieces by the sliding window method for the reading process. The image pieces obtained from the lines are then sent to the text recognition model. First, features are obtained by passing the incoming image pieces through the encoder consisting of Convolution layers. Then, these features are passed through the Bidirectional Long-Short Term Memory (Bi-LSTM) and Connectionist Temporal Classification (CTC) layer, which is the decoder part of the model, and text outputs are produced.

Unlike these [1], [3], [6] studies, [7] study describes a new method for reading clipped words from document data. In this method, the incoming word image is first passed through an encoder called the Gated Recurrent Convolution Neural Network (GRCNN). The features obtained in the encoder section are then sent to the decoder consisting of Bi-LSTM and CTC layers, and text outputs are produced. The aforementioned [1], [3], [6], [7] studies were developed using document data, similar to our study. In the OCR field, there is also OCR in Natural Scene data, in addition to document data. Although the OCR issue in Natural Scene data is similar to the OCR issue in document data, it is a much more difficult problem. While some studies on Natural Scene data focused only on text recognition [8], [9], [10], some others [11], [12], [13], [14] tried to solve the text detection and recognition issue together.

## III. METHODOLOGY

In this section, all stages of the developed OCR engine are examined in detail. The OCR engine mentioned in this study was developed through three basic processes. In the preliminary preparation stage, which is the first stage; the incoming receipt image is subjected to multiple processes for the detection of text fields. In the second stage, word positions are determined on the full-screen receipt image, which is the output of the first stage. In the final stage, the detected words are clipped and sent to the "word reading model" and the reading process is performed. The diagram given in Fig. 1 summarizes the process mentioned. The details of the processes are given in the following subheadings.

### A. Preliminary Preparation

Receipt images sent to the OCR engine and expected to be read often come with background images and taken from different angles. The input image can also contain a background image, as shown in Fig. 2. In such a case, the detection of text fields is even more difficult. Factors such as the constant change of the background, the position of the receipt in the image, the obliquity of the receipt are just some of the reasons that will make it difficult to determine the text field. For this reason, the receipt image is passed through the preliminary preparation stage before the text fields on it are detected. The purpose of the preliminary stage is to obtain a full-screen and vertical image of the receipt in the image, free from the background image, and to send the clean receipt image to the text detection stage.

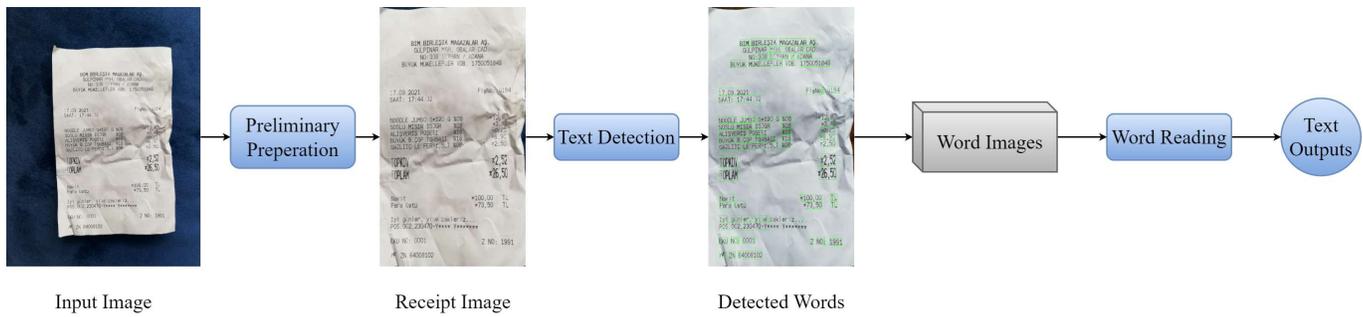


Fig. 1. The diagram that shows all the stages and outputs of the OCR model

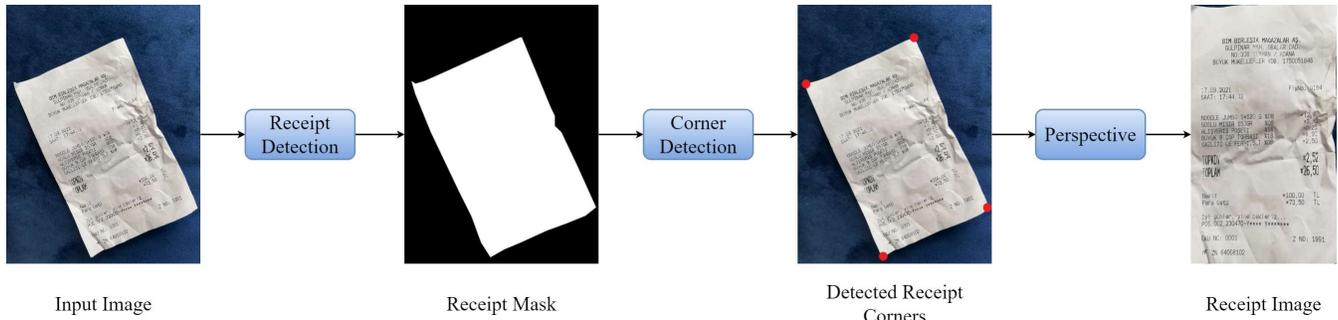


Fig. 2. The processes and outputs of the preliminary stage

Fig. 2 also appears to detail the inside of the box called Pre-processing in Fig. 1. The preliminary preparation phase consists of three steps. These can be listed as (i) determination of the receipt area (receipt mask formation), (ii) determination of receipt corners, and (iii) application of the perspective process. Masking model is used to determine the receipt area. The purpose of the masking model is to produce a mask of the objects in the visual and labeled in the training data. In the training of the masking model, the input visual for the input and the mask of the same resolution as the expected output are required. An example of an input image is the Input Image in Fig. 2. While the mask of this image is produced, the receipt field in the image is labeled with the help of a tool called LabelMe [15]. The tagged receipt field has a value of 1, while the background without a receipt is 0. When this matrix is converted to a image, the Receipt Mask given in Fig. 2 is obtained. This is how the dataset for the training of the masking model is prepared.

The masking model is based on the Convolutional Encoder-Decoder method. The resolution of the model input is reduced as the layers progress. This part is known as the Encoder and is the part where the Model draws information from the image. Then, in the Decoder stage, the image that is reduced in the Encoder stage is restored again. Finally, the mask of the input image is produced by passing through the classification layer. Depending on the number of classes in the training dataset, the activation function of this classification layer can be Sigmoid or Softmax. Since the OCR engine mentioned in this study has a single class of receipt field, the last layer of the masking model used is passed through the Sigmoid activation function. As an output, a matrix consisting of numbers between 0 and 1 in the input image resolution is obtained. In this matrix, pixels

that are close to zero are considered as backgrounds, while pixels with a value closer to one are considered as receipt areas. The receipt area in the receipt image is thus estimated.

There are multiple Encoder-Decoder methods for the masking model. As the masking model used in this study, the most appropriate Encoder-Decoder method was selected among the methods presented in the segmentation\_models [16] library by trying. Among the methods available in this library, only the achievements of Unet [17] and Linknet [18] Encoder-Decoder methods are compared. After selecting the Encoder-Decoder method for the masking model, it is necessary to select a Backbone for the Encoder. As skeletal model, VGG16 [19] and Resnet34 [20] models with similar parameter numbers are compared. A mobilenetv2 [21] model with a lower number of parameters has also been added to the comparison. As a result, the most appropriate masking model was selected by comparing two different Encoder-Decoder methods and three different Skeletons. The results obtained are shown in Table I.

When comparing, dataset and training parameters were kept constant. The dataset used in the comparison consists of the same data as the dataset of the masking model used in the OCR engine mentioned in this study. The dataset contains a

TABLE I. COMPARISON OF COMPARING THE SUCCESSES OF MASKING MODELS. (BATCH SIZE: 8, EPOCHS: 50, METRICS: IOU, SAVEBESTVAL\_IOU\_ACC, INPUT RESOLUTION: 512,512)

Model	Mobilenetv2 Unet	VGG16 Unet	Resnet34 Unet	Mobilenetv2 Linknet	VGG16 Linknet	Resnet34 Linknet
Duration (Sec)	0.0926	0.1070	0.0908	0.0880	0.1004	<b>0.0870</b>
IoU%	98.28	98.42	98.40	98.17	<b>98.44</b>	98.31

total of 1352 background receipt images. 10 percent of this data is reserved for testing. The results given in Table I were obtained using this training and testing dataset. In training, bce\_jaccard\_loss was used as a loss function. The remaining parameters are found in the description of Table I. Each model is trained by up to 50 epochs and the best Val\_IoU values are used for success comparison. Then the average elapsed time for a receipt estimate was used to compare the model speed. Test data were used for duration and success measurement. When the results given in Table I are examined, it is understood that there is not much difference between them. However, if it is necessary to choose the most suitable model, the Resnet34 Unet should be chosen due to its proximity to the highest success and lowest speed. In this study, Resnet34 is used as Skeleton and Unet is used as Encoder-Decoder method in the masking model.

Using the masking model, the receipt areas in the incoming receipt image are determined. This process alone is not enough. The detected receipt area needs to be separated from the unnecessary background. Perspective process is used in this study both to get rid of unnecessary background and to correct oblique receipt images. The perspective process converts images that are oblique, such as the input image given in Fig. 2, into full-screen images such as the Receipt Image in Fig. 2. For this, the four corners of the region to be applied to the perspective process (in this case, the receipt area in the image) must be determined correctly. The receipt mask obtained from the Masking model is used for corner determination. The edges of the receipt image are determined using the Canny function in the OpenCV library on the receipt mask. Then, the output consisting of black and white images with edges is sent to the findContours function and the corner points in the receipt mask are determined. Using the positions of the obtained corner points, the best four corners to represent the receipt mask are selected. Finally, the positions of these four designated verticals are sent to the getPerspectiveTransform. The output of this function is sent to the warpPerspective function and the Receipt Image in Fig. 2 with perspective applied is generated.

**B. Text Detection**

At this stage, the receipt visual, which is the output of the preliminary preparation stage, is taken as input. In order to read the incoming receipt image, the text fields must first be determined. In this study, object detection algorithm is used for the detection of text fields. Object detection algorithms are used to determine the position of pre-labeled objects in the image or video. In this study, the task of the object detection algorithm is to detect the text fields on the receipt image that comes as a full screen. The dataset used for the object detection algorithm is labeled using a tool called LabelMe. The data used in the dataset are the receipt images from the preliminary preparation stage. When this data is labeled, each word in the receipt image is labeled as belonging to the same object class. The dataset consists entirely of Turkish receipt images and consists of 552 data in total. 10 percent of this dataset is reserved for test data.

In this study, You Only Look Once (YOLO) model is used due to its speed and success. For word detection, YOLO V3 [22] and YOLO V4 [23] models were compared and the most appropriate model was selected. For the training, a dataset



Fig. 3. The Words Detected by YOLO Models on the Sample Turkish Receipt (Blue Boxes), the Output of the YOLO V3 (Left), the Output of the YOLO V4 (Right)

consisting of Turkish receipt data used in Text Detection was used. For success comparison, 10 percent of the dataset was used as a test dataset. Accuracy was taken into account as a benchmark. In comparison, the batch size and model resolution are kept the same. Average Precision, IoU, F1 Score are used as comparison metrics. The darknet [24] library, which is also used in YOLO training, was used to calculate the mentioned metrics of YOLO models. The results from the comparison are shared in Table II. In Fig. 3, the words detected by both YOLO models in the sample Turkish receipt image are seen. Although the difference in success is not clearly seen in Fig. 3, when Table II is examined, it is seen that the YOLO V4 model gives the best result in all categories. For this reason, YOLO V4 model is used in the word detection model.

**C. Word Reading**

The word reading stage is reached with the word positions determined during the text detection phase. At this stage, words are clipped from the receipt image using word positions. The cropped words are read by means of the reading model. C-RNN is used as a word reading model in this study. The dataset used for the training of the model consists of words clipped from Turkish receipt images. Words are manually tagged. Although Turkish data are used in the training dataset, the words are labeled according to the Latin alphabet. The reason for this is that there are letters in Turkish that are very similar to each other (ç-c, ö-o, i-ı ...). These letters create ambiguity due to their similarities, and labeling them as a single class increases success by eliminating uncertainty. For this reason, the training dataset is labeled according to the characters in the Latin alphabet. The dataset consists of 68,000 Turkish words

TABLE II. COMPARISON OF THE ACHIEVEMENTS OF YOLO MODELS

Model	Average Precision	IoU	F1 Score
YOLO V3	84.31	59.50	0.80
YOLO V4	<b>86.61</b>	<b>65.5</b>	<b>0.83</b>

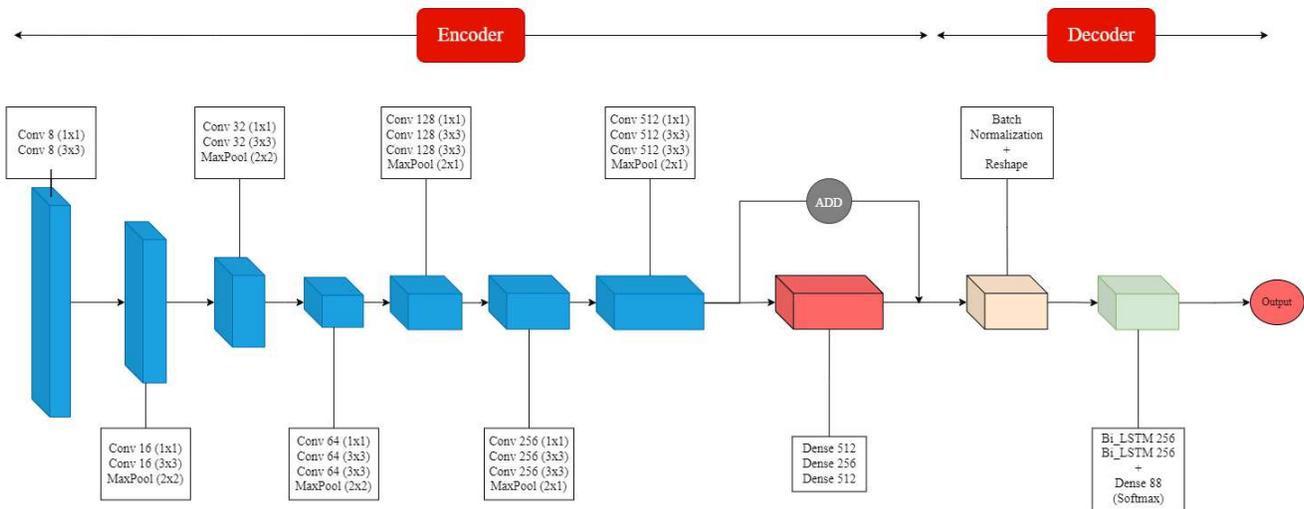


Fig. 4. The C-RNN model scheme used as a word reading model

that are clipped and labeled from approximately 1,100 receipt images. 10 percent of this Turkish data is reserved for the test dataset.

The C-RNN model used for the word reading model is based on the principle of Convolutional Encoder and Recurrent Decoder. As seen in Fig. 4, the incoming word image is first passed through the Convolutional layers. In this section, the information of the letters in the word is obtained. Then, wordprintouts are produced by looking at this information in the Bi-LSTM layer. CTC-Loss is used as the loss function in the model. The model takes words with resolution (256,64,1) as input. During the model prediction, word predictions are produced by passing the word outputs through the ctc\_decode function in the [25] Keras library. After this stage, the words on the receipt and the positions of these words are known. Using this information, the lines are determined. Thus, the text output with specified lines that are the output of the OCR engine is produced.

The word reading model consists of two parts: Encoder and Decoder. While creating the word reading model seen in Fig. 4, various methods were tried on the Encoder and Decoder layers. If we need to talk about the methods used in the Encoder section, some of them can be listed as follows: (i) The stride value was made 2 by using the Convolution layer instead of the pooling layer. (ii) If there are three consecutive Convolution layers, the first layer output is summed up with the third layer output. (iii) The number of Convolution layers before the pooling layer was reduced to one in order to create a simpler model and to increase success. (iv) The activation function used and the output dimension were changed. (v) A feed forward layer has been added to the end of the convolution layer, as shown in Fig. 4. Multiple encoder models were developed using such methods, and the most successful Encoder model seen in Fig. 4 was selected among them. When changing the encoder model, the parameters used in the training and the Decoder model were kept completely constant.

After the encoder model is selected, the Decoder model should be selected as the most successful. The methods used

for this can be listed as follows: (i) The activation function and output dimension used were changed. (ii) Instead of using bidirectional LSTM, two forward directional and two backward directional LSTMs were used and their outputs were combined. (iii) It was attempted to transfer the LSTM hidden state to the next LSTM layer, but this method was not used because it reduced the success as expected. (iv) LSTM outputs were collected with each other. (v) A feed forward layer has been added after the LSTM output. While trying these methods, the training parameters and the encoder model were kept constant. Among the methods tried, the Decoder model in Fig. 4 was selected that gave the most successful result.

## IV. EVALUATIONS

### A. Comparison Dataset

While the success of the OCR engine mentioned in this study was measured, the accuracy of the outputs obtained on the same dataset was compared with other OCR engines. The comparison was performed both in the dataset consisting of Turkish receipts and in the dataset consisting of English receipts. For the success comparison, the production of the OCR engine mentioned in this study requires a dataset that is not used in the preliminary preparation, text detection or word reading stages. A test dataset consisting of Turkish receipts was created for comparison. This dataset consists of 66 Turkish receipt images. These receipt data are labeled on a per-word basis. In Fig. 4, a sample image from the Turkish test dataset is given. When the receipt image is examined, there are Turkish letters such as "İ, Ş, Ü, Ç" in it. Since the OCR engine developed could only read the Latin alphabet, these Turkish letters were converted to similar letters in the Latin alphabet ("I, S, U, C") and tagged. For the sake of equality in comparison, the outputs of all compared OCR engines are also translated into the Latin alphabet. This process was applied only in the Turkish dataset.

As mentioned before, the Turkish dataset used in the comparison cannot be shared because it contains personal information. For this reason, in addition to the Turkish dataset,

the ICDAR2019-SROIE dataset, which consists of English receipt data, is also used. Since the OCR engines were also compared in the receipt dataset in the ICDAR2019-SROIE competition, test images of the same dataset were used to compare the OCR engine produced in this study with other OCR engines. Since the text positions are given on a word-by-word basis in the outputs of the compared OCR engines, the dataset to be compared must also be labeled as words. Since the ICDAR2019-SROIE dataset where the comparison will be made is labeled on a sentence-by-sentence basis, these tags need to be broken down and converted into word tags. When performing this operation, if there are spaces in the labeled text in the dataset, this text is divided into more words than the number of spaces. The process of shredding this tagged text was carried out by us. This English dataset is labeled by the contest holders with all letters capitalized. For this reason, the comparison was made after the outputs of all OCR engines were converted to capital letters. This process was applied to the English dataset only.

The application called LabelMe was used when labeling words. Thanks to the LabelMe application, words and word positions are saved in a file by manually labeling the data. Each word saved in the file is saved in the (LeftTopCornerX LeftTopCornerY RightLowerCornerX RightLowerCornerY Word) format. Only a single word is stored in each line of the file and only the words in a receipt are stored in each file. The order in which words are labeled is not important because then each tagged word is matched to the words read by OCR engines.

### B. Comparison Method

The comparison of OCR engines is based on speed and accuracy. The accuracy category includes two main issues. These are the success of Word Reading and Word Position Detection. These categories are the same as task 1 and task 2 in the ICDAR2019-SROIE study. The first task in the competition measures how accurately the contestants determine the positions of the words on the receipt. In the second task of the competition, it is measured how accurately the competitors read the text in the receipt image. In the ICDAR2019-SROIE competition, they used Precision, Recall and F1-Score to measure success in both tasks. These metrics are not enough to measure the success of an OCR engine. For this reason, while comparing OCR engines in this study, in addition to these metrics, Character Error Rate (CER) metric was used in word reading success and IoU metric was used in word position determination. As mentioned in the Comparison Dataset section of the article each receipt is labeled mixed on a word basis. Since the words and positions obtained by reading the receipt image with the OCR engine are also mixed, it is necessary to match the labeled data with the data read by the OCR engine to measure the success of the OCR engine.

When matching words, it is necessary to match them correctly. OCR engines may have guessed more or less than the words tagged. In order to prevent this situation, it is aimed to achieve the highest success in matching. Compared the tagged words with all the predicted words and recorded CER achievements. The achievements are recorded in the list and converted into a table in the size of (Number of Tagged Word X Number of Words Read). Then, starting from the first row of the table, the word with the highest achievement in the row

is determined. If this word also has the highest achievement in the column and is higher than the specified threshold value, the labeled word in the index of the row it is in is matched to the read word in the index of the column in which it is located. To prevent the matched words from being re-matched, the achievement values in the table where CER achievements are recorded are equalized to zero. Calculations are made after all the words are matched to each other. When performing calculations, the paired words are taken as True Positive ( $TP$ ), while the predicted but unmatched words are taken as False Positive. In addition, words that are not predicted but are labeled are considered False Negative ( $FN$ ). The total number of words matched in a receipt is shown as  $TP_n$ . With these definitions, formulas to be used in success comparison are produced. The formulas used are shown as Eq. (1) and Eq. (2).

## V. RESULTS

Two different categories are taken into account when comparing OCR engines. These are Speed and Accuracy. The accuracy category is similar to the ICDAR2019-SROIE competition and is divided into two as Word Position Detection Success and Word Reading Success. The speed of OCR engines depends on the system being tested. In particular, offline OCR engines such as Easy OCR and Tesseract OCR depend on the power of the system. For this reason, all OCR engines were tested from the same computer and success measurements were made. The processor of the system in which the tests are carried out is Intel i5-11400H 2.7 GHz, the amount of Ram is 16GB, the Graphics Card is NVIDIA RTX 3050TI (Mobile) and CUDA Version 11.7.

### A. Speed Achievement

For companies that use the OCR engine in a product, the document reading speed of the OCR engine is important. The use of a slow OCR engine in a project involving the use of the OCR engine leads to the accumulation of work, thus wasting time for the customer, the project user, and therefore dissatisfaction. For this reason, the speeds of OCR engines are compared in this section. When measuring the speed of the OCR engine, the reading speeds of all the receipts in the dataset used were measured. The sum of these speeds is then averaged by dividing them by the total number of receipts. This results in the average time spent by the OCR engine on a receipt. While the speed of Easy OCR, Tesseract OCR and Nacsoft OCR depends on the system specifications, the speed cannot be measured on the computer where the test results are obtained because the Google Vision Api does not work on the local computer. For this reason, Google Vision Api is not included in the speed measurement. The results obtained in Table III are given.

TABLE III. COMPARISON OF THE DURATION ACHIEVEMENTS OF OCR ENGINES IN THE ENGLISH DATASET

OCR Engine	Easy OCR	Tesseract OCR	Nacsoft OCR
Mean of Time (sec)	0.89	1.88	<b>0.45</b>

TABLE IV. COMPARISON OF THE WORD READING ACHIEVEMENTS OF OCR ENGINES IN THE TURKISH DATASET

OCR Engine	Easy OCR	Tesseract OCR	Vision Api	Nacsoft OCR
Word Reading Success (CER)	79.49	85.10	91.48	<b>93.89</b>
Precision	71.57	82.67	84.12	<b>88.04</b>
Recall	78.26	39.88	<b>92.71</b>	90.07
F1 Score	74.66	52.68	88.15	<b>89.01</b>

TABLE V. COMPARISON OF THE WORD READING ACHIEVEMENTS OF OCR ENGINES IN THE ENGLISH DATASET

OCR Engine	Easy OCR	Tesseract OCR	Vision Api	Nacsoft OCR
Word Reading Success (CER)	88.50	92.90	<b>94.63</b>	90.77
Precision	84.90	<b>92.58</b>	86.30	83.92
Recall	83.11	75.80	<b>92.41</b>	80.91
F1 Score	83.75	82.72	<b>89.13</b>	81.88

### B. Word Reading Success

Another important metric of OCR engines is Word Reading Success. The accuracy of the information contained in the document read by the OCR engine directly depends on the success of reading words. The success of reading words also affects the success in extracting information from the text read. Although there is no OCR engine that reads all documents correctly, it is desired to choose the OCR engine that gives the highest success possible. In this section, the success of reading the words on the receipts is compared. When measuring reading success, the word matching method mentioned in the Method section was used and then the CER value was measured among the matched words. The achievements are added together and divided by the total number of matched words ( $TP_n$ ). Thus, the Word Reading Success of a receipt as defined in Eq. (1) is measured. The achievements of these receipts are then summed up and divided by the number of receipts in the dataset. As a result, the Word Reading Success of the dataset is obtained. Table IV shows the success results in the Turkish receipt dataset, while Table V shows the success results obtained from the English receipt dataset.

$$\frac{\sum_{i=1}^{TP_n} CER_i}{TP_n} \quad (1)$$

When Table IV and Table V are examined, Tesseract OCR has high precision and low recall accuracy. This means that the Tesseract OCR engine predicts a small fraction of the words it needs to guess, but the words it predicts are mostly the words that are on the receipt. This difference between the precision and recall categories is only visible in the Tesseract OCR engine. When the results in Table IV are examined, it is seen that the Nacsoft OCR engine, whose methods are described in this study, has the best accuracy rate in almost all categories. When Table V is examined, it is seen that Nacsoft OCR engine gives close results to other OCR engines. Judging by the results of the Vision api in Table IV and Table V, it is seen that it gives the best success in the dataset consisting of English receipts, and in the Turkish dataset, it comes after the Nacsoft OCR engine with a close difference.

TABLE VI. COMPARISON OF THE WORD POSITION DETECTION ACHIEVEMENTS OF OCR ENGINES IN THE TURKISH DATASET

OCR Engine	Easy OCR	Tesseract OCR	Vision Api	Nacsoft OCR
Word Position Detection Success (IoU)	65.34	77.04	<b>84.03</b>	78.75
Precision	85.55	80.02	88.82	<b>92.68</b>
Recall	93.37	38.85	<b>97.90</b>	94.79
F1 Score	89.16	51.26	93.08	<b>93.69</b>

TABLE VII. COMPARISON OF THE WORD POSITION DETECTION ACHIEVEMENTS OF OCR ENGINES IN THE ENGLISH DATASET

OCR Engine	Easy OCR	Tesseract OCR	Vision Api	Nacsoft OCR
Word Position Detection Success (IoU)	66.87	<b>77.73</b>	75.94	76.86
Precision	92.10	<b>93.13</b>	90.01	90.74
Recall	89.92	76.44	<b>96.38</b>	86.88
F1 Score	90.72	83.40	<b>92.96</b>	88.16

### C. Word Position Detection Success

Determining the position of the word is important in extracting information from the text. IoU is used when calculating the success of word positioning. IoU is measured between the matched words in the word matching section mentioned in the method section. By adding up the achievements and dividing them by the total number of matched words, the Word Position Determination success of a receipt data as defined in Eq. (2) is measured. In this way, the IoU success of all receipts in the dataset is measured and the results obtained are collected. Then, the total value obtained is divided by the number of receipts in the dataset, measuring the Word Position Detection Success of the OCR engine in the dataset. Table VI shows the success results in the Turkish receipt dataset, while Table VII shows the success results obtained from the English receipt dataset.

$$\frac{\sum_{i=1}^{TP_n} IoU_i}{TP_n} \quad (2)$$

An examination of the results of the Tesseract OCR engine in Table VI and Table VII shows high precision and low recall as mentioned earlier. When the results of the Easy OCR engine are examined, Table VI and Table VII have the lowest accuracy of word position detection success. The Vision Api has the best success in both Turkish and English datasets. A review of the results of the Nacsoft OCR engine in Table VI shows that it has the best F1-Score accuracy compared to other OCR engines, but it surpasses the Vision Api in word position detection achievement.

## VI. CONCLUSION

In the study, OCR engine was developed on Turkish receipt data by using artificial intelligence methods. The success of the developed Nacsoft OCR engine was compared with the success of Tesseract OCR, Easy OCR and Google Vision Api on English and Turkish receipt data. When the results obtained are examined, Nacsoft OCR gives better results in Turkish receipt data than other open source OCR engines, but cannot

reach the same result in English receipt data. This may be due to the fact that the training dataset consists only of Turkish receipts. In addition, the low number of data used in the training of the Nacsoft OCR engine may also adversely affect success. According to this situation, success can be increased by increasing the number of data used in the training of the Nacsoft OCR engine and adding English data to the training dataset. In the training of the Nacsoft OCR engine in future studies, other document types besides the receipt data can be included in the training.

#### ACKNOWLEDGMENTS

We would like to thank Nacsoft for providing their anonymous receipt data.

#### DATA AVAILABILITY

The Turkish datasets used in this study to train the models cannot be shared due to they contain personal data. English receipt dataset that used for comparison in our study and also in ICDAR-2019-SROIE competition can be found here.

#### CONFLICT OF INTEREST

We have no conflicts of interest to disclose.

#### FUNDING DECLARATION

This work is supported by Turkey's National Small and Medium Enterprises Development Organization (KOSGEB). The project's title is "The development of machine learning software enabling the reading, analysis, and administration of invoices and receipts".

#### REFERENCES

- [1] R. Smith, "An Overview of the Tesseract OCR Engine," in *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*, vol. 2, Sep. 2007, pp. 629–633, iSSN: 2379-2140.
- [2] "EasyOCR," Sep. 2022, original-date: 2020-03-14T11:46:39Z. [Online]. Available: <https://github.com/JaidedAI/EasyOCR>
- [3] T. M. Breuel, "The OCRopus Open Source OCR System."
- [4] Z. Huang, K. Chen, J. He, X. Bai, D. Karatzas, S. Lu, and C. V. Jawahar, "ICDAR2019 Competition on Scanned Receipt OCR and Information Extraction," in *2019 International Conference on Document Analysis and Recognition (ICDAR)*, Sep. 2019, pp. 1516–1520, arXiv:2103.10213 [cs]. [Online]. Available: <http://arxiv.org/abs/2103.10213>
- [5] B. Shi, X. Bai, and C. Yao, "An End-to-End Trainable Neural Network for Image-based Sequence Recognition and Its Application to Scene Text Recognition," Jul. 2015, arXiv:1507.05717 [cs]. [Online]. Available: <http://arxiv.org/abs/1507.05717>
- [6] H. El Bahi and A. Zatni, "Text recognition in document images obtained by a smartphone based on deep convolutional and recurrent neural network," *Multimedia Tools and Applications*, vol. 78, no. 18, pp. 26453–26481, Sep. 2019. [Online]. Available: <https://doi.org/10.1007/s11042-019-07855-z>
- [7] J. Wang and X. Hu, "Gated Recurrent Convolution Neural Network for OCR," in *Advances in Neural Information Processing Systems*, vol. 30. Curran Associates, Inc., 2017. [Online]. Available: <https://proceedings.neurips.cc/paper/2017/hash/c24cd76e1ce41366a4bbe8a49b02a028-Abstract.html>
- [8] R. Atienza, "Vision Transformer for Fast and Efficient Scene Text Recognition," May 2021, arXiv:2105.08582 [cs]. [Online]. Available: <http://arxiv.org/abs/2105.08582>
- [9] O. Alsharif and J. Pineau, "End-to-End Text Recognition with Hybrid HMM Maxout Models," Oct. 2013, arXiv:1310.1811 [cs]. [Online]. Available: <http://arxiv.org/abs/1310.1811>
- [10] X. Tang, Y. Lai, Y. Liu, Y. Fu, and R. Fang, "Visual-Semantic Transformer for Scene Text Recognition," Dec. 2021, arXiv:2112.00948 [cs]. [Online]. Available: <http://arxiv.org/abs/2112.00948>
- [11] K. Wang, B. Babenko, and S. Belongie, "End-to-end scene text recognition," in *2011 International Conference on Computer Vision*, Nov. 2011, pp. 1457–1464, iSSN: 2380-7504.
- [12] M. Busta, L. Neumann, and J. Matas, "Deep TextSpotter: An End-to-End Trainable Scene Text Localization and Recognition Framework," in *2017 IEEE International Conference on Computer Vision (ICCV)*. Venice: IEEE, Oct. 2017, pp. 2223–2231. [Online]. Available: <http://ieeexplore.ieee.org/document/8237504/>
- [13] T. He, Z. Tian, W. Huang, C. Shen, Y. Qiao, and C. Sun, "An End-to-End TextSpotter with Explicit Alignment and Attention," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Salt Lake City, UT, USA: IEEE, Jun. 2018, pp. 5020–5029. [Online]. Available: <https://ieeexplore.ieee.org/document/8578625/>
- [14] Y. Liu, H. Chen, C. Shen, T. He, L. Jin, and L. Wang, "ABCNet: Real-Time Scene Text Spotting With Adaptive Bezier-Curve Network," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Seattle, WA, USA: IEEE, Jun. 2020, pp. 9806–9815. [Online]. Available: <https://ieeexplore.ieee.org/document/9156344/>
- [15] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman, "LabelMe: A Database and Web-Based Tool for Image Annotation," *International Journal of Computer Vision*, vol. 77, no. 1-3, pp. 157–173, May 2008. [Online]. Available: <http://link.springer.com/10.1007/s11263-007-0090-8>
- [16] P. Iakubovskii, "qubvel/segmentation\_models," Sep. 2022, original-date: 2018-06-05T13:27:56Z. [Online]. Available: [https://github.com/qubvel/segmentation\\_models](https://github.com/qubvel/segmentation_models)
- [17] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," May 2015, arXiv:1505.04597 [cs]. [Online]. Available: <http://arxiv.org/abs/1505.04597>
- [18] A. Chaurasia and E. Culurciello, "LinkNet: Exploiting Encoder Representations for Efficient Semantic Segmentation," in *2017 IEEE Visual Communications and Image Processing (VCIP)*, Dec. 2017, pp. 1–4, arXiv:1707.03718 [cs]. [Online]. Available: <http://arxiv.org/abs/1707.03718>
- [19] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," Apr. 2015, arXiv:1409.1556 [cs]. [Online]. Available: <http://arxiv.org/abs/1409.1556>
- [20] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," Dec. 2015, arXiv:1512.03385 [cs]. [Online]. Available: <http://arxiv.org/abs/1512.03385>
- [21] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted Residuals and Linear Bottlenecks," Mar. 2019, arXiv:1801.04381 [cs]. [Online]. Available: <http://arxiv.org/abs/1801.04381>
- [22] J. Redmon and A. Farhadi, "YOLOv3: An Incremental Improvement," Apr. 2018, arXiv:1804.02767 [cs]. [Online]. Available: <http://arxiv.org/abs/1804.02767>
- [23] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOv4: Optimal Speed and Accuracy of Object Detection," Apr. 2020, arXiv:2004.10934 [cs, eess]. [Online]. Available: <http://arxiv.org/abs/2004.10934>
- [24] Alexey, "Yolo v4, v3 and v2 for Windows and Linux," Oct. 2022, original-date: 2016-12-02T11:14:00Z. [Online]. Available: <https://github.com/AlexeyAB/darknet>
- [25] "Keras: Deep Learning for humans," Sep. 2022, original-date: 2015-03-28T00:35:42Z. [Online]. Available: <https://github.com/keras-team/keras>

# A Survey on Blockchain Technology Concepts, Applications and Security

Asma Mubark Alqahtani<sup>1</sup>, Abdulmohsen Algarni<sup>2</sup>  
College of Computer Science, King Khalid University  
Abha, Asir, Saudi Arabia<sup>1,2</sup>

**Abstract**—In the past decade, blockchain technology has become increasingly prevalent in our daily lives. This technology consists of a chain of blocks that contains the history of transactions and information about its users. Distributed digital ledgers are used in blockchain. A transparent environment is created by using this technology, allowing encrypted secure transactions to be verified and approved by all users. As a powerful tool, blockchain can be utilized for a wide range of useful purposes in everyday life including cryptocurrency, Internet-of-Things (IoT), finance, reputation system, and healthcare. This paper aims to provide an overview of blockchain technology and its security issues for users and researchers. In particular, those who conduct their business using blockchain technology. This paper includes a comparison of consensus algorithms and a description of cryptography. Further, most applications used in blockchain are focused on in this paper also analyzing real attacks and then summarizing security measures in blockchain. Even though Blockchain holds a promising scope of development in several sectors, it is prone to several security and vulnerability issues that arise from different types of blockchain networks which represent a challenge to deal with blockchain. Finally, as a research community, we encourage future research challenges that can be addressed to improve security in blockchain systems.

**Keywords**—Blockchain; cryptography; cryptocurrency; consensus algorithms; blockchain security

## I. INTRODUCTION

Blockchain is based on a decentralized, unchangeable database that makes it simpler to record assets and keep track of transactions in a corporate network. An asset may be tangible or intangible. On a blockchain network, virtually anything of value may be stored and traded, reducing risk and improving efficiency for all users. Generally, a blockchain is a digital ledger of transactions that are being recorded. It is decentralized and is not controlled by any individual, group, or company [1].

As a structured technology, blockchain can be very difficult to change without the approval of the people who use it. Blockchain stores data as a decentralized ledger. Participants in this network can read, write, and verify transactions. Transactions cannot be modified or deleted. To support and secure the blockchain system, digital signatures, hash functions, and other cryptographic functions are used. These primitives ensure that transactions recorded in the ledger are integrity-protected and authenticated. This technology is called blockchain because new blocks are linked to older ones to form a chain. The first appearance of this term was a publication written by S. Haber and W.S. Stornetta in 1991 [2]. In general, blockchain technology is credited to Satoshi Nakamoto, who developed

the theory and implemented the technology in 2008 and 2009, respectively in the cryptocurrency Bitcoin, the most well-known blockchain application. Blockchain technology in recent years has attracted significant attention from academics and industries because of its advanced features. It can be applied to a variety of applications beyond cryptocurrencies. Blockchain technology has become a leading technology of internet interaction systems, including the Internet of Things (IoT) [3].

Our motivation in this paper is to inform and assist someone to become familiar with blockchain technology and its security issues, particularly for those who carry out transactions using blockchain technology and for researchers interested in developing blockchain technology and evaluating its security issues. To search publications and information on the Internet, the first step is to identify keywords such as blockchain, consensus algorithm, cryptography, cryptocurrency, and blockchain security. A second approach is to review papers that have been published in top conferences and journals that deal with blockchain. In this paper, we provide the following main contributions:

- A detailed survey was conducted on blockchain technology.
- A systematic survey of Blockchain applications is conducted in this paper. 10 application areas are considered.
- Security and privacy issues were also addressed.

Therefore, we encourage further efforts to survey and develop blockchain technology for widespread adoption.

The rest of this paper consists of the following sections: In Section II, we provide an overview of the history of blockchain technology. A typical consensus algorithm used in the blockchain is described in Section III. In Section IV, we focused on blockchain applications. In Section V, we summarize the technical risks, attacks, and challenges of security in this area, and in Section VI, we conclude this paper.

## II. HISTORY OF BLOCKCHAIN

Chaum's Ph.D. thesis, published in 1982, was the first to suggest a blockchain as a protocol. A paper by Haber and Stornetta published in 1991 titled "How to Time-Stamp a Digital Document" detailed the concept of time stamping digital data cryptographically [3].

In 1998, Nick Szabo proposed the creation of Bit Gold, an early attempt at the creation of a decentralized virtual currency.

However, Szabo’s attempt to implement Bit Gold is generally regarded as the basis for Satoshi Nakamoto’s bitcoin protocol, even though the project was never implemented [4].

Modern day blockchain technology is widely believed to have been first implemented by Satoshi Nakamoto in 2008. He hypothesized a direct online payment between parties without the use of a third-party intermediary. Rather than relying on trust, that paper presented a cryptographic proof-based electronic payment system [5].

Blockchain was introduced by Ethereum in 2013 as a technology for executing smart contracts on a decentralized platform. With Ethereum, it is possible for developers to create markets, store transactions, and move funds according to written instructions, all without the involvement of middlemen. Unlike Bitcoin, Ethereum is a ledger technology that is being used by companies to develop new programs, which are being expanded beyond the realm of currencies for the first time [6]. With the launch of the Ethereum platform in 2015, blockchain could be used for storing and processing loans and contacts. Using an algorithm known as a smart contract, this technology ensures the implementation of an action between two parties. Due to Ethereum’s ability to provide a faster, safer, and more efficient environment, it became extremely popular. Instead of all the different blockchain projects, Ethereum enables communication via untrusted distributed applications on its own blockchain, thus creating a new concept called Ethereum 2.0 [7].

Hyperledger is open source software for blockchains that was announced by the Linux Foundation in 2015. The Hyperledger blockchain framework aims to build enterprise blockchains, which are different from Bitcoin and Ethereum. Blockchain attracted interest with its capability to enable anonymity, but the real appeal lies in its capability to enable complete privacy. As will be discussed in the fourth section, there have been many applications for blockchain technology that have been discovered across a wide range of industries.

The following Fig. 1 summarizes the history of blockchain technology. Since everyone can participate in Bitcoin and Ethereum’s blockchain networks, they are considered public blockchains. Due to their need to verify participants before joining the network, the Hyperledger blockchain networks are considered private blockchains, also known as permissioned blockchains. The following Table I summarizes the differences between Hyperledger and Ethereum, two popular blockchain platforms and networks.

TABLE I. HYPERLEDGER AND ETHEREUM

Feature	Ethereum	Hyperledger
Purpose	run smart contracts	Businesses
Confidentiality	public network	limited access
Governance	Ethereum developers	Linux Foundation
Participation	permission-free	Only authorized members
Smart contracts	Yes	Yes using chaincode
Programming Language	Solidity	JavaScript, Java, etc
Consensus Mechanism	Yes, PoW,PoS, etc	No
Speed of Transactions	Low	High
Use	Public Applications	Private applications

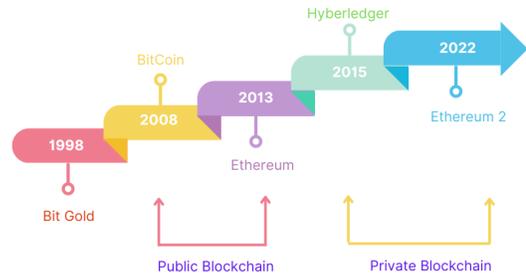


Fig. 1. History of blockchain.

### III. BLOCKCHAIN TECHNOLOGY

#### A. Blockchain Layers

According to Melanie Swan, blockchain technology has passed through two stages. The first stage is blockchain 1.0 represented by Bitcoin, and the second stage is blockchain 2.0 represented by Ethereum. In general, blockchain-based technologies include Bitcoin, Ethereum, Hyperledger, etc [8]. Even though the implementations are varied, there are some similarities in the basic architecture.

Blockchain environments can be classified into five layers, as shown in Table II application, network, contract, consensus, and data layers.

TABLE II. BLOCKCHAIN LAYERS

Layers	Bitcoin	Ethereum	Hyperledger
Application layer	Bitcoin trading	Ethereum trading	Enterprise blockchain
Network layer	TCP	TCP	HTTP
Contract layer	Script	Script	Java
Consensus layer	PoW	PoW/PoS	PBFT/SBFT
Data layer	Merkle tree	Merkle patricia tree	Merkle Bocket tree

Consensus mechanisms are the main component of the consensus layer. In the contract layer, smart contracts are included. Various protocols for data transmission and verification are included in the network layer. In addition, it is pertinent to note that the blockchain is a typical peer-to-peer network. There is no central node and all nodes are connected through a planar topology [9]. It is possible to transact between any two nodes. Each node within the network is free to leave or join anytime. A number of applications are included in the application layer, such as Bitcoin, Ethereum, and Hyperledger.

#### B. Consensus Algorithms

Among the many desirable characteristics of blockchain technology, it is possible to verify the honesty of anonymous users when they enter transactions into the ledger. This is done by validating each transaction to ensure that it is legal

before adding it to a block. Consensus algorithms are used to determine whether new blocks will be added to the blockchain and to ensure trust between parties involved in the blockchain system and to store transactions. As a result, consensus algorithms are the core of all blockchain transactions [10]. Every participant must follow a consensus protocol. There have been several consensus mechanisms developed for blockchains. This includes Proof of State, Delegated Proof of State, Proof of Work, Proof of Elapsed Time, Directed Acyclic Graph, and so on. We will take a look at the most common algorithms shown in Table III.

**Proof of Work (PoW):** The objective of this algorithm is to determine a problem that must be solved through guessing. Bitcoin and Ethereum employ PoW as the algorithm for their consensus. As a result of PoW requiring lots of electricity and time, it is not widely used [11].

**Proof of Stake (PoS):** It ranks second in popularity as a consensus algorithm, and it involves fewer computations than PoW. It minimizes the time and energy waste issues that PoW has. This consensus algorithm replaces the current method for reaching consensus in a distributed system, instead of solving a Proof-of-Work. BlackCoin was the first cryptocurrency to use a PoS [12].

**Proof of Elapsed Time (PoET):** It is a consensus algorithm for blockchain networks that keeps the process more efficient by avoiding over-utilization of resources and high-energy consumption. The PoET method resembles the proof of work method (PoW), but requires less power due to its ability to allow the processor to switch to other tasks after a period of time, which increases efficiency [13].

**Byzantine Fault Tolerance (BFT):** It is aimed at solving problems where there are untrustworthy parties, but they need to achieve consensus. PBFT is designed to improve BFT. With PBFT, if hostile nodes represent fewer than thirty percent of all nodes, then the current state of the blockchain will be agreed upon by all participants. Blockchain systems are more secure when there are more nodes involved. Currently, Hyperledger Fabric is based on PBFT [14].

**Direct Acyclic Graph (DAG):** It consists of vertices and edges, which differentiates it from various consensus algorithms. Transactions are represented by the vertices of the structure. A block is not referred to in this algorithm, nor do we need to use a mining process to add transactions. Each transaction is built upon the previous one rather than being grouped into a block. Several applications of DAG technology can be found in fields that require high speed and no fees, like Internet of Things (IoT) [15].

TABLE III. CONSENSUS ALGORITHMS

Algorithms	Speed	setup	Example of use
PoW	Low	Public/private	Bitcoin, Ethereum
PoS	High	Public/private	NXT
DPoS	High	Public/private	EOS, BitShares
DBFT	Very high	Public/private	NEO, TON
PBFT	High	Private	Hyperledger, Chain

### C. Smart Contract

The smart contract also called chaincode is an essential feature of blockchain because it not only offers a distributed, immutable completion of all activities, but is also capable of allowing for the creation of a computer program that is non-subjective and specifies how the process will be implemented. In this contract, an important activity is addressed. More than two parties don't need to be involved in this contract. The Ethereum smart contract was designed to overcome some of the limitations of Bitcoin [16].

Enterprise blockchain applications are based on smart contracts, which will revolutionize the way businesses operate. Smart contracts can be developed by anyone without the need for an intermediary. Because of a smart contract, the process is autonomous, accurate, and cost-effective.

### D. Cryptography of Blockchain

Blockchains enable confidential and secure transactions between anonymous parties. This trust is established through cryptography, thus eliminating the necessity for centralized institutions. By using cryptography, blockchain data is kept on the ledger. Cryptography building blocks are used in blockchain technology as follows [17]:

- **Public Key Cryptography:** Designed to create digital signatures and encrypt data.
- **Zero-Knowledge Proof:** Show that you know a secret without divulging it.
- **Hash Functions:** A mathematical function that generates pseudo-random numbers.

*1) Public key cryptography:* A transaction can be proven to have been created by the right user by this method. Using a private key, a user can sign a message, known as a digital signature. Digital signatures are used in Hyperledger and Ethereum transactions to verify the authenticity of the sender and that the information has not been changed since it was signed. The algorithm (ECDSA) is widely used to generate a combined set of private and public keys.

*2) Zero-knowledge proofs:* These are primarily used when users request to transfer money to other users. Before committing a transaction, the blockchain must verify that the participant who is transferring funds has enough to complete the transaction. However, the blockchain does not care about how much money he has in total or who is spending it so it has no idea who the user is or how much money he owns.

*3) Hash functions:* Hash Functions: Hash functions form an essential part of blockchain technology. There are five properties of a hash function that are critical for cryptography [18]:

**Fixed size:** The hash function can accept any input and create the output of a fixed size. In order to provide digital signatures, blockchains employ hash functions to condense messages.

**Preimage resistance:** When given a set of inputs, it is not challenging to produce a hash result. Despite this, reverse engineering the original input is mathematically impossible

based on the hash output. The only way to achieve the same result is to randomly select data that should be entered into the hash algorithm.

2nd preimage resistance: Obtaining a secondary input that provides the same hash result is impossible given an input and its hash result.

**Collision resistance:** The same hash output cannot be produced from two distinct inputs.

**Big change:** An entirely different hash output will be produced if any single bit is changed in the input.

#### IV. BLOCKCHAIN APPLICATIONS

According to the survey, blockchain applications include cryptocurrency, Internet-of-Things (IoT), finance, reputation system, healthcare, security and privacy, advertising, copyright protection, society application, energy, mobile applications, defense, digital records, supply chain, digital ownership management, automotive, intrusion detection, agricultural sector, voting, identity management, education, law and enforcement, property title registries, asset tracking, and so on [19]. An illustration of the spiraling applications of blockchain can be found in Fig. 2.

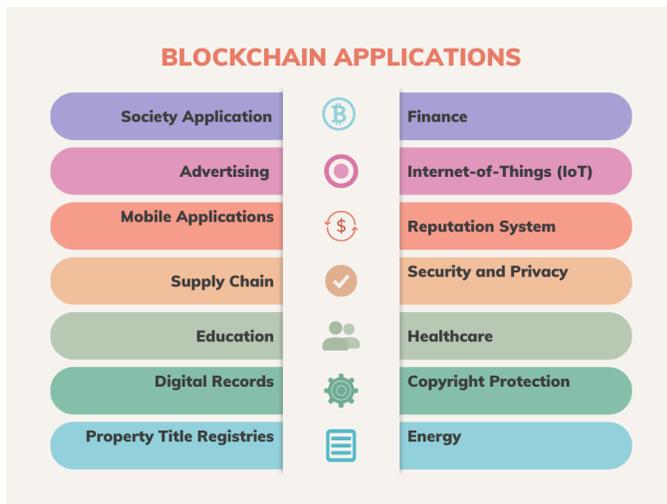


Fig. 2. Application of blockchain.

More applications of blockchain systems are predicted to be developed in the future. To provide further information, we have selected the following 10 blockchain-based applications:

##### A. Healthcare

Prescription medications are being tracked and traced throughout supply chains using blockchain technology. The tool enables the easy and rapid prevention and regulation of counterfeit pharmaceutical distribution as well as the recall of ineffective and unsafe medications. Security of customer data is a primary goal in healthcare, as is the exchange of data between hospitals, governments, and research institutes, which facilitates the improvement of healthcare services. As part of this project, Nokia has used wearable devices to track daily steps and hours of sleep and stored the data on the Blockchain [20].

##### B. IoT

People, places, and products can be connected via the Internet of Things (IoT), providing new opportunities for the generation of value in products and business processes. On the other hand, implementing this technology on a large scale is fraught with security concerns. Combining blockchain and IoT offers the following benefits: To detect data manipulation quickly and accurately, blockchain technology can provide a robust framework for faster detection. Due to the size of IoT networks, it can be difficult to detect failure patterns. Each IoT endpoint is assigned a unique key by blockchain technology, which facilitates the identification of inconsistencies. By combining IoT with smart contracts, it becomes possible to authorize automated responses. Decentralization enhances security: Blockchain technology is decentralized, making it impossible for cybercriminals to hack and corrupt a single server. Additionally, the use of blockchain technology allows tracking of user actions to provide information on who, when, and how users have used a particular device [21].

##### C. Government

Blockchain technology can be used in the public sector to improve the quality and quantity of services. It can also be used to improve transparency and accessibility, as well as to share information between different organizations. In addition to being secure against online attacks, the blockchain is publicly available. Transactions are not editable or deletable once they have been added. This makes data transactions safe, secure, and accessible to anyone [22].

##### D. Power Grid

The development of blockchain-based smart grids is aimed at improving energy distribution on a large scale. There is a significant amount of inefficiency in electricity distribution at the retail level. The use of blockchain technology and Internet-of-Things (IoT) devices for these types of services can reduce electricity bills by bypassing retailers and directly connecting consumers to wholesale distributors. Consumers connected to the smart grid can also shop around for the highest rates from a variety of providers. This leveled the playing field in an industry that has traditionally been dominated by a single provider. Several projects are leading the way in this area, including Grid + and Energy Web Token [23].

##### E. Copyright and Royalties

Music, films, and other creative mediums are subject to copyright and royalties. These are artistic mediums and do not appear to be linked to Blockchain in any way. In the creative industries, however, this technology is quite critical in terms of ensuring security and transparency. It is common for music, films, art, etc., to be plagiarized without proper credit being given to the original creators. A detailed ledger of artist rights can be maintained on the Blockchain to rectify this issue. The use of blockchain technology can also provide a secure record of artist royalties and deals with large production companies, in addition to being transparent. Digital currencies, such as Bitcoin, can also be used to manage the payment of royalties [24].

### F. Cryptocurrencies

In 2008, it was announced that Bitcoin would be the first cryptocurrency. It was launched in 2009. It is estimated that there are 21 million bitcoins in use today. The miner receives a transaction fee once he finds a value that matches the difficulty. Currently, about 90% of BTC is mined.

Ethereum (ETH) is regarded as the second largest cryptocurrency based on market capitalization after Bitcoin (BTC). According to Cryptoslate, [25] there are 2403 top cryptocurrencies ranked by market capitalization. Table IV below shows seven popular cryptocurrencies.

TABLE IV. TOP CRYPTOCURRENCIES

No	cryptocurrency	Market cap
1	Bitcoin (BTC)	322.5 billion
2	Ethereum (ETH)	162.8 billion
3	Tether (USDT)	66.3 billion
4	Binance Coin (BNB)	44.0 billion
5	U.S. Dollar Coin (USDC)	43.9 billion
6	XRP (XRP)	17.66 billion
7	Binance USD (BUSD)	16.3 billion

Blockchain technology can be applied to the use of cryptocurrencies, thus taking full advantage of the features of this technology including:

- There is no intermediary involved in the payment process.
- Processing fees are low.
- Money can be sent at any time without delay or restriction.

A few disadvantages of cryptocurrencies include:

- Black money may be incurred due to a lack of control.
- Digital assets may be lost as a result of a security attack, which we will discuss in more detail later.
- Some commentators claim that investing in cryptocurrencies is highly speculative and risky. Tesla, for instance, advised investors to be aware of Bitcoin's volatility.

### G. Dubai Blockchain Office

Strategy of Dubai Blockchain is the result of a collaboration between the Dubai Future Foundation and the Digital Dubai Office. The purpose of this initiative is to continuously explore and evaluate the latest technological innovations that can be used to enhance the quality of life in cities through seamless, efficient, safe, and impactful solutions [26].

The strategy represents a powerful and innovative tool to influence the future of the Internet through the provision of safe and simple transactions. This will help to achieve the vision of making Dubai the world's first blockchain-powered city. When this strategy is successful, Dubai will contribute substantially to the future economy.

### H. Cloud Computing

Cloud computing has had a major impact on the software technology industry due to its impressive benefits. There are many uses for cloud computing among businesses worldwide, including data storage and backup, software development and testing, disaster recovery, and more. Many industries are using cloud computing to build innovative solutions, including healthcare, automotive, and retail. Even with the advantages of cloud computing, it has its limitations. Blockchain can help overcome these limitations. Due to its transparency, security, and decentralized nature, blockchain technology is being used by millions of businesses for a variety of industrial applications. The use of blockchain and cloud technology together, however, can further revolutionize industries. Even though blockchain technology provides better network security, privacy, and decentralization, cloud computing provides high scalability and elasticity. Therefore, cloud technology and blockchain technology can be combined to produce innovative solutions [27].

### I. e-Commerce

Constant evolution is taking place in the e-commerce industry due to the development of new technologies and the creation of new ways to buy and sell products and services. Using blockchain technology, it is possible to create a decentralized database for storing information about products and customers. By doing so, customers would be able to obtain information about products, such as their origin and supplier, which would also reduce the possibility of fraud. A blockchain-based payment system can also ensure enhanced security and reduce the risk of fraudulent payments. As a distributed database, blockchain technology provides secure, transparent, and tamper-proof transactions. It is anticipated that this technology will revolutionize the e-commerce industry by improving the security of transactions and simplifying the fulfillment process. The system also enhances buyer-seller trust and transparency. Blockchain technology allows e-commerce businesses to track the history of orders and transactions to improve the customer experience. The customer would be able to track their orders easier and find information about previous purchases. Additionally, blockchain can reduce the risk of fraud and facilitate the tracking and verification of transactions more reliably and securely. The implementation of this technology could prove to be a game changer for the e-commerce industry, which is currently plagued by issues of fake reviews, fraudulent transactions, and other security risks. Businesses that use blockchain technology can reduce costs associated with processing transactions and shipping products, as well as improve the speed at which new products are introduced to the market [28].

### J. Advertising

A blockchain advertising application is a type of distributed ledger technology that promotes decentralization with the highest level of security and transparency. On the blockchain, digital records are immutable, which means that individuals have access to read but cannot amend the records. Blockchain can allow advertisers to track their advertising expenditures in real time since it stores information and transactions. It provides a level of transparency that cannot be achieved with

existing systems. Transparency is not the only advantage. In advertising, speed is crucial, as it is difficult to track inventory and ensure high-quality products. Blockchain technology has the capability of keeping up with these challenges [29].

## V. ATTACKS AND SECURITY MEASURE ON BLOCKCHAIN

### A. Attacks on Blockchains

Blockchains are distributed so it makes sense to conduct research on their security. In this section, we will discuss the security risks associated with this technology. In order to gain a deeper understanding of blockchain security, it is essential to first understand the differences between private and public blockchain security, particularly regarding data access and participation capabilities, as we mentioned above.

The following are the top security issues associated with blockchains [30]:

1) *Sybil attack*: In this attack, several fake network nodes are generated by hackers. Through the use of these nodes, it will be able to gain majority agreement and interrupt transactions.

2) *Endpoint vulnerabilities*: Another vital concern in the security of blockchain is the vulnerability of endpoints. Electronic devices such as mobile phones and computers are used to interact with the blockchain network. Observing the behavior of users and targeting their devices will allow hackers to steal the user's key. Perhaps this is one of the most prominent security issues associated with blockchain technology.

3) *51% attack*: An attack of 51% occurs when one user or institution controls half of the hash rate and takes control of the entire system. Transactions can be modified by hackers and prevent them from being confirmed. They will even reverse transactions that have already been completed, leading to double spending.

4) *Phishing attacks*: Phishing attacks are designed to steal user credentials. An email will be sent to the wallet key owner that appears to be legitimate. A fake hyperlink is attached to the email that requires the user to enter their login details. By gaining access to a user's credentials and private information, it is possible to cause damage to the user and the blockchain network as a whole.

5) *Routing attacks*: In this attack, participants are usually unaware of the threat because the transmission of data and the conduct of operations continue as usual. A potential danger is that such attacks could reveal sensitive information or generate revenue without the user's permission. There is a critical reliance on the movement in real time of enormous amounts of information in a blockchain application and network. Due to the anonymity of an account, hackers may be able to intercept information transmitted to Internet service providers by using it.

6) *Private keys*: You will need a private key in order to access your funds. A hacker can easily guess the private key if it is weak. Your funds could be accessed as a result. Keeping your private key secret is extremely critical, and it should be strong enough not to be guessed easily.

7) *Malicious nodes*: Additional security problems related to blockchain technology include the threat of malicious nodes. An attempt to disrupt the network will occur once a dishonest actor has joined the network. In order to accomplish this, they will attempt to reverse transactions or flood the network with transactions.

### B. Security Measures of Blockchain

To ensure the security of blockchain applications, security must be considered at all layers, including permission management through several security measures [31]. The following are some of the security measures of blockchain:

1) *Blockchain governance*: Determining how existing organizations or users leave or join the network, and providing mechanisms to prevent malicious actors, manage errors, secure data, and address issues between parties.

2) *Data security*: While data compression is generally regarded as the most effective method for identifying what data should be kept on-chain, additional privacy measures should be implemented to hash data, cloud storage, and data in transit.

3) *Security of blockchain network*: Blockchain is a distributed system, which requires network connections from various participants beyond a single organization to interact. All of these factors have the potential to introduce security exploits or flaws. Part of governance, therefore, includes reviewing security protocols for users [32].

4) *Blockchain application security*: Security applications are vulnerable points and should be protected with effective user identification and endpoint security measures. For private blockchains, where access and use are limited to authorized participants, it may be necessary to provide different levels of authorization that may change with time.

5) *Smart Contracts Security*: Smart contracts consist of a set of codes within the blockchain, triggered by a set of programmed conditions. This presents another point of vulnerability as their reliability determines whether the operation and the results can be trusted.

6) *Use of trusted third-parties*: Security evaluations, penetration checks, and reviews of the source code of smart contracts and blockchain implementations should be performed only by trusted individuals. Use these to protect against new security threats, such as unauthorized access to cryptographic algorithms [33].

## VI. CONCLUSION

During the past few years, blockchain technology has attracted a great deal of attention due to its advanced characteristics of decentralization, autonomy, integrity, immutability, verification, and fault tolerance. In terms of the future scope, the primary priority will be addressing the security concerns arising from the various types of blockchain networks. Furthermore, consensus algorithms such as PoW implemented on blockchain have several drawbacks. Thus, the development of a consensus algorithm that is more efficient will result in more cost-effective blockchain networks. This survey introduces an in-depth overview of blockchain technology. A brief historical overview of blockchain was presented, followed by

a comparison of the most widely used consensus algorithms. It has been discussed in detail how public key cryptography and hash functions applied to blockchains can be used for security, identification, and non-repudiation purposes. In addition, it provides detailed information and comparisons of some cryptocurrencies used in blockchain. Also, we focus on various categories of top security risks associated with blockchain technology. Finally, by making this effort, we hope that someone will gain a deeper understanding of blockchain technology. We also hope that individuals will give more focus to the safety of the blockchain

#### ACKNOWLEDGMENT

This research is financially supported by the Deanship of Scientific Research at King Khalid University under research grant number (R.G.P.1/188/41).

#### REFERENCES

- [1] Kumar, S., Kumar, A., and Verma, V. (2019). A survey paper on blockchain technology, challenges and opportunities. *Int. J. Comput. Trends Technol.(IJCTT)*, 67(4), 16. ISO 690
- [2] Haber, S., & Stornetta, W. S. (1991). How to time-stamp a digital document (pp. 437-455). Springer Berlin Heidelberg.
- [3] Zheng, Z., Xie, S., Dai, H. N., Chen, X., and Wang, H. (2018). Blockchain challenges and opportunities: A survey. *International journal of web and grid services*, 14(4), 352-375. ISO 690
- [4] R. Sharma, Bit gold, Investopedia, 2021. Available online: <https://www.investopedia.com/terms/b/bit-gold.asp>.
- [5] S. Nakamoto, Bitcoin: a peer-to-peer electronic cash system. <https://bitcoin.org/bitcoin.pdf>, October 2008.
- [6] Vujičić, D., Jagodić, D., & Randić, S. (2018, March). Blockchain technology, bitcoin, and Ethereum: A brief overview. In 2018 17th international symposium infotech-jahorina (infotech) (pp. 1-6). IEEE.
- [7] A. Groetsema, A. Groetsema, N. Sahdev, N. Salami, R. Schwentker, F. Cioanca, Blockchain for Business: an Introduction to Hyperledger Technologies, The Linux Foundation, 2019.
- [8] Sarmah, S. S. (2018). Understanding blockchain technology. *Computer Science and Engineering*, 8(2), 23-29. ISO 690
- [9] Zhai, S., Yang, Y., Li, J., Qiu, C., & Zhao, J. (2019, February). Research on the Application of Cryptography on the Blockchain. In *Journal of Physics: Conference Series* (Vol. 1168, No. 3, p. 032077). IOP Publishing.
- [10] Chaudhry, N., and Yousaf, M. M. (2018, December). Consensus algorithms in blockchain: comparative analysis, challenges and opportunities. In 2018 12th International Conference on Open Source Systems and Technologies (ICOSST) (pp. 54-63). IEEE. ISO 690
- [11] Nguyen, G. T., and Kim, K. (2018). A survey about consensus algorithms used in blockchain. *Journal of Information processing systems*, 14(1), 101-128. ISO 690.
- [12] Saad, S. M. S., & Radzi, R. Z. R. M. (2020). Comparative review of the blockchain consensus algorithm between proof of stake (pos) and delegated proof of stake (dpos). *International Journal of Innovative Computing*, 10(2).
- [13] J. Frankenfield, Proof of Elapsed Time (PoET) (Cryptocurrency), Invest, October 16, 2020. Available online: <https://www.investopedia.com/terms/p/proof-elapsed>
- [14] Zhang, Z., Zhu, D., & Fan, W. (2020, December). Qpbft: practical byzantine fault tolerance consensus algorithm based on quantified-role. In 2020 IEEE 19th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom) (pp. 991-997). IEEE. ISO 690
- [15] Ferdous, M. S., Chowdhury, M. J. M., Hoque, M. A., & Colman, A. (2020). Blockchain consensus algorithms: A survey. arXiv preprint arXiv:2001.07091. ISO 690
- [16] Khan, S.N., Loukil, F., Ghedira-Guegan, C. et al. Blockchain smart contracts: Applications, challenges, and future trends. *Peer-to-Peer Netw. Appl.* 14, 2901–2925 (2021). <https://doi.org/10.1007/s12083-021-01127-0>
- [17] R. Santos, K. Bennett, E. Lee, Blockchain: Understanding its Uses and Implications, The Linux Foundation, 2021. Available online: <https://www.edx.org/course/blockchain-understanding-its-uses-and-implications>.
- [18] Wang, M., Duan, M., & Zhu, J. (2018, May). Research on the security criteria of hash functions in the blockchain. In *Proceedings of the 2nd ACM Workshop on Blockchains, Cryptocurrencies, and Contracts* (pp. 47-55). ISO 690
- [19] Ali, M. S., Vecchio, M., Pincheira, M., Dolui, K., Antonelli, F., and Rehmani, M. H. (2018). Applications of blockchains in the Internet of Things: A comprehensive survey. *IEEE Communications Surveys and Tutorials*, 21(2), 1676-1717.
- [20] Agbo, C. C., Mahmoud, Q. H., & Eklund, J. M. (2019, April). Blockchain technology in healthcare: a systematic review. In *Healthcare* (Vol. 7, No. 2, p. 56). MDPI.
- [21] Panarello, A., Tapas, N., Merlino, G., Longo, F., & Puliafito, A. (2018). Blockchain and IoT integration: A systematic survey. *Sensors*, 18(8), 2575. ISO 690
- [22] Alketbi, A., Nasir, Q., & Talib, M. A. (2018, February). Blockchain for government services—Use cases, security benefits, and challenges. In 2018 15th Learning and Technology Conference (L&T) (pp. 112-119). IEEE.
- [23] Foti, M., & Vavalis, M. (2021). What blockchain can do for power grids? *Blockchain: Research and Applications*, 2(1), 100008.
- [24] Kim, A., & Kim, M. (2020, October). A study on blockchain-based music distribution framework: focusing on copyright protection. In 2020 International conference on information and communication technology convergence (ICTC) (pp. 1921-1925). IEEE.
- [25] Aggarwal, S., & Kumar, N. (2021). History of blockchain-Blockchain 1.0: Currency. In *Advances in Computers* (Vol. 121, pp. 147-169). Elsevier. ISO 690
- [26] Avan-Nomayo, O. Dubai's Economic Department to Roll Out Blockchain-Based Corporate KYC. Available online: <https://cointelegraph.com/news/dubai-s-economic-department-to-roll-out-blockchain-based-corporate-kyc>.
- [27] Gai, K., Guo, J., Zhu, L., & Yu, S. (2020). Blockchain meets cloud computing: A survey. *IEEE Communications Surveys & Tutorials*, 22(3), 2009-2030.
- [28] Zhou, Z., Wang, M., Yang, C. N., Fu, Z., Sun, X., & Wu, Q. J. (2021). Blockchain-based decentralized reputation system in an E-commerce environment. *Future Generation Computer Systems*, 124, 155-167.
- [29] Chen, W., Xu, Z., Shi, S., Zhao, Y., & Zhao, J. (2018, December). A survey of blockchain applications in different domains. In *Proceedings of the 2018 International Conference on Blockchain Technology and Application* (pp. 17-21).
- [30] Aruba, 10 Blockchain and New Age Security Attacks You Should Know, January 22, 2019. Available online: <https://blogs.arubanetworks.com/solutions/10-blockchain-and-new-age-security-attacks-you-should-know/>.
- [31] Guo, H., & Yu, X. (2022). A Survey on Blockchain Technology and its security. *Blockchain: Research and Applications*, 3(2), 100067.
- [32] D. Wang, J. Zhao and Y. Wang, "A Survey on Privacy Protection of Blockchain: The Technology and Application," in *IEEE Access*, vol. 8, pp. 108766-108781, 2020, doi: 10.1109/ACCESS.2020.2994294.
- [33] Velmurugadass, P., Dhanasekaran, S., Anand, S. S., & Vasudevan, V. (2021). Enhancing Blockchain security in cloud computing with IoT environment using ECIES and cryptography hash algorithm. *Materials Today: Proceedings*, 37, 2653-2659. ISO 690

# An Autonomous Role and Consideration of Electronic Health Systems with Access Control in Developed Countries: A Review

Mohd Rafiz Salji<sup>1</sup>, Nur Izura Udzir<sup>2</sup>

Faculty of Information Management, Universiti Teknologi MARA,  
Malaysia<sup>1</sup>

Faculty of Computer Science and Information Technology, Universiti Putra  
Malaysia<sup>2</sup>

**Abstract**—The electronic healthcare system (EHS) nowadays is essential to access, maintain, store, and share the electronic health records (EHR) of patients. It should provide safer, more efficient, and cost-effective healthcare. There are several challenges with EHS, notably in terms of security and privacy. Nonetheless, many approaches can be utilized to tackle it, and one of them is access control. Even though numerous access control models were presented, traditional methods of access control, such as role-based access control (RBAC), were extensively employed and are still in use today. Currently, the number of EHS equipped with access control keeps growing, and some previous works utilize RBAC only or an autonomous role. However, relying only on a role in today's advanced technology may jeopardize security and privacy. The previous work also has flaws because of using an ineffective instrument that is costly to maintain and will burden organizations, particularly in developed countries. In this paper, the background and emphasis on the challenges associated with an autonomous role in the EHS are discussed. Following that, this paper provides recommendations and analytical discussion on existing EHSs with access control mechanisms for securing and protecting EHR in developed countries. Finally, instrument information in the form of a SWOT analysis is recommended to replace the present instrument utilized by the previous work for a notion to the organizations in the developed countries to select the best environment for their future or upgrade EHS.

**Keywords**—Access control; security; privacy; electronic health-care system; electronic health record; developed countries

## I. INTRODUCTION

Recently, developments in information technology have made significant progress in the field of medical information. This information nowadays is managed in a system called EHS. Advanced EHS is required to manage massive volumes of EHR clearly and cost-effectively. EHR is a non-printed form that electronically stores all of a patient's medical information. EHS is now developed and administered by numerous medical institution systems that allow sharing of EHR among various healthcare practitioners and organizations, rather than by a single healthcare organization. Due to that, a sophisticated system must be put in place to secure and preserve EHR.

Many approaches have been utilized to address security and privacy protection, but the most commonly used is access control. In general, access control is important in securing systems and protecting the privacy of authorized users, especially when providing health services. Today, many types of EHS

combined with access control have been proposed, however, there are still issues that impede the development of EHS with access control.

This study discusses the issue of the utilization of access control mechanisms in EHS. The current EHS [1] has been established to secure and preserve EHR in developed countries. With regards to security and privacy, the previous system was in the initial stage of discussing the plan to use some instruments to secure their system, and the EHR in this system was protected using RBAC, which uses role or job function to allow or deny access to resources. Subsequently, another previous system [2], also used the RBAC model to secure the storage server in the cloud environment. Based on these previous works [1], [2], an autonomous role was utilized in these systems. It is indisputable that this traditional access control model is still applicable today; however, the problem is that employing an autonomous role to secure systems and preserve privacy in today's advances of ICT is not acceptable and unsafe, especially in a healthcare environment, due to various drawbacks of this scheme observed by previous works. Details about RBAC are discussed in Section II. The previous work [1] also suffers from the problem of utilizing the wrong choice of instrument, since they utilize a centralized database that has been claimed to have issues such as timely, frequent errors, and costly [3], [4].

This article review's main contributions are grouped into four categories. First, based on the problems of the previous works [1], [2], this paper provides background information on the RBAC model, including a description of the model's history, followed by examining the scheme's benefits and drawbacks based on previous works, and finally, discussions on the current EHSs employing autonomous RBAC. Second, this article examines and suggests previous literature reviews related to current EHS with access control models by discovering the environment and mechanisms used by previous systems instead of relying solely on an anonymous role. The aim is as a notion or an opportunity for the clinics and hospitals in developed countries to enhance their EHS with appropriate mechanisms in a specific environment in the future. It also aids researchers in swiftly grasping each function of the mechanisms employed by prior systems. The current EHS with access control models is examined in three categories: EHS with access controls aimed at securing the system, existing systems aimed at protecting the EHR, and finally,

existing systems capable of securing the system and protecting the EHR. Third, based on previous literature reviews, this article provides an analytical discussion in terms of the issues or problems of previous works, findings, or results of the previous works, and finally, comments or suggestions based on previous works. This discussion aimed to help identify how many issues or problems were faced by previous works and group them in the same categories. This discussion also helps clinics and hospitals in developed countries in obtaining information about the problems that happen in EHS and the way to solve them and have a notion to develop or upgrade EHS. It also helps the researcher in understanding difficulties and the ideas of previous works in solving problems, and comments or suggestions can be used to find opportunities for future work. Finally, because of previous work employed the wrong instrument, this paper provides critical instrument selection information in the form of a SWOT analysis that can assist clinics and hospitals in developed countries choose an acceptable and cost-effective instrument.

The rest of this paper is organized as follows: Section II provides the background of the RBAC model, while consideration of EHS with access control is discussed in Section III. In Section IV, the SWOT analysis is presented, and finally, Section V concludes the work.

## II. ROLE-BASED ACCESS CONTROL (RBAC)

Role-Based Access Control (RBAC) means to allow or deny client data from being accessed by the user based on role, i.e., job function or position, however, this decision depends on organizational policy [5]. This model has been introduced for over twenty years, primarily in UNIX, and centralized computer conditions, yet this model needs standardization because every framework utilizes its restrictive elements [6]. Therefore, the National Institute of Standards and Technology (NIST) began a task in 1992 to bound together with the principles of RBAC by incorporating the current models [6]. Although RBAC has since quite a while ago existed and is viewed as traditional access control, this model is still being used and stays pertinent right up until today.

In a positive sense, there are quite a few points why RBAC has become well known and can be utilized by current systems. The advantages of RBAC are as follows:

- 1) Simplifies access management and user permission review [7], where it is easy to categorize roles and a group of users for each role [8] and it aids in determining which permissions are permitted for which users in a large enterprise system [9].
- 2) RBAC policies adhere to the need-to-know security concept and fulfil the notion of least-access privileges [10]. This model also may be well-known for managing complicated role hierarchies in organizations [11].
- 3) This model may not need to be concerned about users being added or deleted from the system because this architecture is ideally suited to a large organization [12].
- 4) It can be considered an acceptable model in a healthcare cloud, as it has key strengths such as efficient management of large-scale user permissions, enforcement of need-to-know access controls, simplified

auditing for regulatory compliance, and scalability [8].

Even though RBAC offers advantages, this model likewise experiences a few limitations. The following are RBAC drawbacks:

- 1) This model is incompatible with an open system in which the user is almost likely unknown, and the system recognizes a user solely by roles without knowing the identity and purpose of access —[13].
- 2) Previous works [14], [15] have highlighted that RBAC can lead to privacy disclosure, especially sensitive attributes to unauthorized and untrusted users because of the insufficient and inefficient of this model.
- 3) RBAC is less flexible and responsive because of its static role. As a result, RBAC cannot define granular control over users in certain roles in accessing certain individual objects, which is generally not sufficient for organizations with complex organizational structures, such as collaborative E-healthcare environments [9], [16], [17].
- 4) In a healthcare environment, installing an emergency access mechanism on a static role can pose a high security threat [18], for example, if unauthorized users can have illegal access rights under RBAC, they can easily compromise health records using the emergency access control window because there are no additional control variables to authenticate attacker access.
- 5) Although a previous study [12] has shown that RBAC is suitable for large organizations, however, RBAC is experiencing a role explosion or lack of scalability due to the increasing number of different roles. Furthermore, maintaining all these roles to provide appropriate access rights can be a difficult task [19]. Therefore, RBAC is not advised to be used in cloud computing or in a large system due to the lack of scalability [20], [21], [4].

In light of the previous passages, the aim of featuring the advantages and the drawbacks is to indicate the performance of this model. Despite the fact that the relevance of using RBAC until now was highlighted, nonetheless, this model also has many limitations. Therefore, proposing access control with an autonomous role, in the current context, i.e., in a collaborative system, is extremely hazardous.

Currently, several EHSs utilizing an autonomous role have been proposed. First, previous work [1] proposes a notion of early implementation of the EHS design model in the clinics and hospitals in developed countries, so that they do not miss out on the benefits of building this system rather than paper-based. A typical hospital workflow was defined and utilized in the design process. This study offers a prototype of an EHR web-based system that secures and protects privacy by utilizing RBAC. However, relying solely on RBAC without supporting other features may cause a security and privacy risk. This system also suffers weaknesses when using RBAC, such as static in nature and inflexibility [9], [16], [18], [17], which pose a difficulty if the user needs to treat patients during an emergency situation. A centralized database is an instrument

used in this system to allow access, maintain and store EHR. However, this instrument is not suitable to be utilized in developed countries since it contradicts the goals of generating cost-effective EHS. Next, Li et al. [22] also propose EHS with RBAC model to protect cloud-based outsourced EHRs. They claimed that this model provides an efficient and secure RBAC strategy for securing EHR stored on a storage server, even if the storage server is administered by an untrustworthy third party. This system offers a distinct and more efficient form of fine-grained access control that does not rely on attribute-based encryption (ABE). Only users with roles that adhere to the access policy are permitted to decapsulate. However, in the current circumstances, adopting an autonomous role may put the system in danger.

To summarize, employing an autonomous role to secure and maintain privacy in an internal, external, or collaborative system setting is not viable in today's tough environment. It is agreed that RBAC is still relevant nowadays since it has numerous benefits, however, this model needs support or a hybrid with other features. In the next section, the recommended current EHSs with access control utilizing with or without roles to secure and preserve the EHR is highlighted.

### III. CONSIDERATION OF EHS WITH ACCESS CONTROL MODEL

This section provides information on current EHSs with access control as a reference or notion for organizations in developed countries to developing efficient and effective EHS. The main aim is to highlight and compare the environment and mechanisms applied in the previous works. This section is divided into three sections: 1) The EHS with access control approaches seeks to secure the system, 2) The EHS with access control mechanisms intends to protect the EHR, and 3) The EHS with access control models to secure and protect EHR. This section also provides an analytical discussion of all collections of previous works in terms of the problems or issues, finding or results, and comments or suggestions.

#### A. Security

There are eight EHSs with access control in a cloud environment, and in this section, these systems are discussed.

First, in the cloud-fog computing environment, a searchable personal health records (PHR) framework with fine-grained access control was proposed. PHR is also EHR, however, PHR is controlled, shared, or maintained by patients themselves to support their personal care [23]. This framework was proposed to address the need for local information for a terminal device and the weaknesses of cloud computing [24]. To provide a keyword search function and fine-grained access control, the proposed framework integrates attribute-based encryption (ABE) technology and search encryption (SE) technology. When the keyword index and trapdoor match are successful, the cloud server provider only delivers relevant search results to the user, resulting in a more accurate search. Experiments with simulations demonstrate that the proposed method works well in a cloud-fog scenario. However, the keyword sets are obtained from the actual encrypted file on the cloud, introducing the prospect of a chosen-ciphertext attack. Besides, a novel, fine-grained, and flexible PHRs data access control system for

cloud computing based on encryption was proposed to address the problem of repeated processes in data encryption [2]. The scheme consists of the symmetric key and the ABE layer. The system supports multi-privilege access control for PHRs from multiple patients in the ABE layer. To resolve the problem of repetitive processes, the scheme combines data encryption from different patients, where data is under a single access policy, to reduce encryption and decryption costs. Through implementation and simulation, the proposed scheme shows efficient in terms of time. Moreover, the proposed scheme proved that it was secured based on the security of the CP-ABE scheme. This system ensures data privacy, but, due to computational complexity and scalability concerns, it is unsuitable for health records. Next, the previous work [25] also utilized a CP-ABE based access control for a smart medical system with policy-hiding capabilities that is secure and efficient to overcome the problem of Zhang et al. [26] approach that fails to offer efficient large data storage with leakage resistance. The access control uses hidden access policies to satisfy the medical user's attribute values. A comparison of performance analysis reveals that the suggested system is more efficient than the current scheme. A Secure Healthcare Framework (SecHS) in the cloud using CP-ABE was proposed to provide secure access to health and medical information [27]. Patient data is encrypted under a symmetric encryption scheme and the access policy in CP-ABE is embedded with the ciphertext. The proposed framework was compared with current CP-ABE frameworks, and it demonstrates that SecHS offers greater features for data security. Next, the User Usage Based Encryption (UUBE) diversified access control framework, which usually builds on the searchable encryption technique to secure outsourced data was proposed [28]. In this method, the owner or patient will outsource data to the cloud data center. Data will be encrypted with a multiuser setting and will be stored in the form of ciphertext and finally stored in the database. To search PHR, the user needs to be authenticated by their category of user and institution. After receiving a request from a user, the data center computes the matching encrypted keyword search and returns the relevant outcome. Usage-based encryption is designed for user access and revoke after a specified time. This approach ensures a high level of security for data sharing. If there is misconduct in data access and various attacks by the revocation of the user, the suggested approach proved efficient. However, granular data access cannot be achieved using standard CP-ABE techniques, instead, a multi message CP-ABE is required. Subsequently, to secure cloud storage, a novel system using a hybrid encryption algorithm using Improved Key Scheme of RSA (IKGSR) and Blowfish was proposed [10]. To efficiently retrieve the encrypted data, steganography-based access control was utilized for key sharing via substring indexing and keyword search mechanisms. The findings clearly show that the proposed technique delivers superior security while also retrieving data more efficiently. An expressive and efficient access control method with attribute/user revocation based on the ordered binary decision diagram (OBDD) access structure was proposed to overcome the previous CP-ABE schemes relying on access structures that are either restrictive or cumbersome, resulting in less expressive and efficient [29]. The proposed work establishes attribute groups, which are made up of users who have specific attributes. Each attribute group has its own group key. Version numbers are assigned to user secret keys and ciphertexts to avoid cooperation between

revoked and non-revoked users. When a user's attribute is revoked, a new attribute group key is produced and disseminated to all group members except the revoked user. When there is a change in the attribute group key following an attribute/user revocation, the version number is incremented. The proposed approach was analyzed regarding security and efficiency, and shows that it is secure, expressive, and efficient. Finally, due to the inflexibility of the RBAC, a cloud-based EHR architecture to implement ABAC that employs extensible access control markup language (XACML) was presented [30]. The proposed approach has two stages, after conducting access control on patient records, encryption and digital signatures are applied as an additional security precaution utilizing XML encryption and XML digital signatures to provide more flexible and fine-grained control and minimize the chances of revealing patient private records. A comparison of the security criteria to those utilized in other relevant research was applied and found that the suggested technique was more secure than previous methods. However, encryption in XML requests and responses, on the other hand, is highly expensive for data sharing. Requests and responses are explicitly communicated between legal parties in the first phase and are thus vulnerable to attack.

Subsequently, the previous discussion of EHS with access control models aimed to secure the system is summarized in the form of a comparative analysis. The explanation is shown in Table I.

## B. Privacy

In this section, eight EHSs with access control models that seek to protect EHR are discussed.

First, a privacy-aware relationship semantics-based extensible access control markup language (XACML) access control model was proposed that uses XACML to execute hybrid relationship and ABAC in the hybrid cloud [4]. To enhance multipurpose EHR utilization, the proposed approach offers fine-grained relation-based access control (Rel BAC) with an anonymization technique called Anatomy as it provides quality data utilization. The proposed model delivers and maintains efficient privacy vs utility trade-off. The proposed model was explicitly validated to assess its efficacy regarding privacy-aware electronic health data access and multi-functional usage. The experimental findings demonstrate that access policies based on relationships and EHR anonymization may perform well in terms of access policy response time, and space storage in the proposed model. Next, due to the patient's reluctance to share sensitive data, organizations rely on cloud solutions that employ machine learning models. This article offers a Euclidean L3P-based Multi-Objective Successive Approximation (EMSA) algorithm, efficient measure of privacy in a cloud [31]. Each EHR is divided into common and privacy-related attributes. Privacy-related attributes, such as sensitive information, are subjected to a cryptographic mechanism to produce a key for storage in a cloud environment. Role-based encryption keys are provided here as the fundamental foundation for the storage of sensitive data in cloud environments. In terms of performance, the proposed EMSA was compared with Bat, PUBAT, TPNGS, WOA, and CIC-WOA algorithms based on performance metrics, such as fitness, privacy, and utility.

According to the simulation, the suggested EMSA model has greater privacy values.

A new framework for access control was proposed that protects the privacy of PHR data while a patient is in an emergency [32]. The system proposed uses smart contracts that may limit PHR access permissions in a state of emergency. The smart contract also enables the PHR owner to assign the rules to an employee (a certified medical practitioner) who has the authorization to access the actual data from the PHR, considering the time restriction. The system suggested provides historical audit records that store the history of transactions in an emergency. The proposed framework, based on the experiment, is improved regarding accessibility, privacy, emergency access control, and data auditing in health care systems. A PHR-based blockchain model was proposed to solve the limitation of the blockchain [33]. The proposed model is constructed to provide a tamper-resistant feature utilizing blockchain technology. To protect privacy, proxy re-encryption, and other cryptographic methods are applied. A comprehensive safety analysis reveals that the proposed model can protect the privacy and tamper resistance. The performance study reveals superior overall performance in the proposed model compared to the current literature approach. This work extended [34] by analyzing the system on a variety of user counts and PHR data sizes in a real-world situation.

Permission to access the EHR requires agreement from the patient (data owner), and additional access authorization to be granted by the patient to the healthcare professional is required. A newly built Health Information System (HIS) access decisions flow, guaranteed by RBAC, incorporating patient-centered control was designed [35]. Colored Petri-Networks (CPN) is used as a mimic for RBAC to demonstrate security policy conflicts or restrictions during the access control authorization process. To provide explicit permission for a patient to access their data in a non-offensive access flow, a discretionary access control (DAC) feature was added. Mutual exclusive was designed to consider patient needs for them to permit healthcare providers to access EHR data. Additional information was added to the permission Access Control matrix to ensure privacy is protected and subject to DAC. When compared to prior CPN simulations, a minor modification is proposed to integrate RBAC-aware systems with no significant drawbacks. Subsequently, a novel healthcare access control model named Solution de Gestion Automatisée du Consentement / automated consent management solution (SGAC) was proposed to manage patient consent for accessing their EHR [36]. Because patient preferences and rules may conflict, the SGAC provides a mechanism to handle this issue based on priority, specificity, and modality. Four sorts of characteristics were examined to safeguard patient privacy while providing effective care in life-threatening situations: accessibility, availability, contextuality, and rule effectivity. The verification of SGAC access control rules utilizing two first-order logic model controls, Alloy, and ProB, based on distinct technologies. The results show that SGAC performs better than XACML and that ProB outperforms Alloy by two orders of magnitude thanks to its programmable approach to constraint solving. A formal specification of the system based on the legislation that defines it was proposed to improve the confidence level of the patient towards the system in privacy preservation [37]. This work concentrated on the control and access features

TABLE I. EHS WITH ACCESS CONTROL MODELS IN THE CLOUD ENVIRONMENT AIMS TO SECURE THE SYSTEM

No	Ref.	Mechanism										
		ABE	CP-ABE	UUBE	SE	Sym metric	IKGSR	Blowfish	Stega-nography	OBDD	ABAC	XACML
1.	Sun, 2018	/			/							
2.	Li, 2018		/			/						
3.	Rana, 2020		/									
4.	Satar, 2021		/									
5.	Suresh, 2019			/								
6.	Chinnasamy, 2021						/	/	/			
7.	Edemacu, 2020									/		
8.	Seol, 2018										/	/

of patients' health information. The method used relies on the correct-by-construction Event-B to prove the control and access properties of the system. Finally, traditional approaches like k-anonymity and its derivations frequently overgeneralize, resulting in lower data accuracy. To address this problem, the Semantic Linkage K-Anonymity (SLKA) method was offered, which allows for continuing record linkages [38]. This work demonstrates how SLKA strikes a balance between privacy and utility preservation by detecting risky combinations hidden in data releases.

Subsequently, the previous discussion of EHSs with access control aims to protect privacy is summarized in the form of comparative analysis. The explanation is shown in Table II.

### C. Security and Privacy

In this section, information about recent works on EHSs with access control to secure the system and protect EHR is highlighted.

A secure sharing architecture based on MA-ABE with anonymous authentication outsourcing was proposed to protect the patients' privacy and guarantee that patients may control their PHRs [39]. Before outsourcing, all PHRs are protected using MA-ABE, which overcomes the key hosting problem and achieves fine-grained access control to PHRs. Furthermore, anonymous authentication between the cloud and the user is recommended in order to secure data integrity on the cloud without revealing the user's identity during authentication. The proposed authentication is based on a novel attribute-based online-offline signature. In comparison to previous studies, the suggested approach not only retains encrypted PHRs resistant to collusion assaults and not forged throughout the sharing time, but it also accomplishes privacy preservation, which improves patients' control over their PHRs. Next, some health institutions in the Republic of South Africa have problems protecting HIV patient data because they still use traditional approaches, e.g., paper-based. This work aims to build a cloud-based access control model to share in nine (9) provinces in the South African Republic [40]. This study is based on the acceptance and use of the RBAC model for permission access based on job function, the Access Control List (ACL) contained a list of access control entries (ACE) to identify

trustees and specify access privileges, and Motive Based Access Control (MBAC) models related to data object and motives of seeking them. However, this framework proposes a static model which is not suitable for emergency conditions. Subsequently, CP-ABE also was employed in the proposed scheme to enhance the retrieval capabilities of data based on disease and to solve the inefficiency of RBAC [41]. The proposed scheme can retrieve encrypted EHR based on a specific disease. Furthermore, the scheme ensures user access control and the anonymity of the user or data owner during data retrieval. Moreover, the scheme is resistant to collusion between unauthorized retrievers to access the data. Based on the results of the analysis, the suggested method accomplishes data confidentiality, user anonymity, and collusion resistance. A unique privacy-preserving access control (PPAC) method for electronic health records (EHR) was proposed based on the attribute-based signcryption (ABSC) scheme and the cuckoo filter [42] to solve the issue of security and privacy in sharing EHR. The ciphertext-policy attribute-based signcryption (CP-ABSC) is proposed to ensure fine-grained access control of the EHR data, utilize a cuckoo filter to hide the access policy, and preserve the privacy of EHR owners. Security analysis reveals that the proposed scheme is provably secure. In addition, the performance study reveals that, compared to previous schemes, the suggested scheme achieves low communication and calculation costs, while maintaining the privacy of its owner. However, hiding the AC policy may result in a loss of efficiency. A multi-layer access control (MLAC) model was proposed for building a secure and privacy-preserving EHR system that allows patients to exchange data with stakeholders [43]. In this article, a dual-layer access control model called pseudo-role attribute-based access control (PR-ABAC) was utilized that incorporates attributes with roles for secure sharing of EHR across many contributors. To protect the integrity of patient data, the proposed system also employs the notion of provenance. PASH, a privacy-aware s-health access control system, was introduced based on a large universe CP-ABE with partially concealed access restrictions to solve the problems of conventional CP-ABE [41]. In PASH, access policy attribute values are concealed in encrypted s-health records (SHRs), and only attribute names are exposed. In reality, attribute values contain far more sensitive data than general attribute names.

TABLE II. EHS WITH ACCESS CONTROL MODELS AIMS TO PROTECT PRIVACY

No.	Ref.	Environment		Mechanism									
		Cloud	Blockchain	XACML	ABAC	Rel BAC	Anatomy	EMSA algorithm	CPN	RBAC	SGAC	Event-B	SLKA
1.	Kanwal, 2019	/		/	/	/	/						
2.	Sathya, 2021	/						/					
3.	Rajput, 2021		/										
4.	Thwin, 2019		/										
5.	Junior, 2020								/	/			
6.	Huynh, 2019										/		
7.	Rivera, 2020											/	
8.	Lu, 2018												/

PASH, in particular, implements an efficient SHR decryption test that requires a limited number of bilinear pairings. The attribute universe can be exponentially huge, whereas public parameters are modest and constant in size. According to security analysis, PASH is completely secure in the standard model. PASH is more efficient and expressive than prior systems, based on performance comparisons and experimental data. However, this system lacks revocation. A sensitive and energetic access control (SEAC) was proposed for managing cloud-hosted EHRs and enabling fine-grained access control even in the critical environment to solve problems of the security of the prior system that have threatened the patient’s privacy [44]. The system suggested guarantees that data from a patient are confidential where only authorized users may be permitted to modify or review particular data from the patient. Before submitting to cloud storage, each EHR data is encrypted by the managing authority. The requesting user can receive rights that change permission dynamically based on authentication and context attributes. The security analysis shows that the SEAC mechanism is secure and prevents unwanted access. The findings indicate outstanding compatibility and performance with various configurations and settings. However, keyword searches on encrypted data are not possible using the encryption methods employed. The encryption technique employs bilinear mapping, which has a high computational cost and is impractical for lightweight applications. A hybrid framework called MediTrust was proposed which combines two systems, namely RBAC and ABE, and operates in a semantic database, guaranteeing that patient data are accessible to various access controls [19]. On the provider side, patient data is encrypted before it is outsourced to the cloud server. After download, it is decrypted again at the user. The patient’s general PHR and medical reports are stored separately on another cloud server. CAPTCHA provides the second stage of security control, particularly for security checks, which allows users to connect to MediTrust. The third step of safety control additionally provides for sharing one key with the registered cell phone number of the user and sharing another key with the user’s e-mail id. In MediTrust, the PHR must be decrypted with the combination of the two keys. Furthermore, Amazon AWS EC2 CA was used to validate ABE policies and access control security mechanisms for privacy preservation on PHR. The

results of performance evaluations demonstrate that regarding time complexity and computational overhead, the proposed MediTrust is superior than the prior projects.

A system was proposed by using a technique known as channelling integrated with a smart contract logic script within the network to ensure interoperability of EHR and access control only through the authorization of the patient [45]. The goal of this approach is to provide the entire privacy, integrity, and access control of distributed EHR. Simulated findings show that the proposed solution uses the blockchain to provide absolute transparency and perfect privacy inside a distributed network of sharing EHRs in the medical setting. Next, a blockchain-based architecture was proposed to secure, interoperable and efficient access to patients’ medical records, while protecting the privacy of sensitive data of patients [46]. The proposed framework, named Ancile, uses smart contracts on an Ethereum-based blockchain for enhanced access control, and data obfuscation, as well as advanced cryptographic methods for additional security. However, this work uses six different forms of smart contracts for a proxy re-encryption approach that may incur high computational costs. Finally, Smart Contract-based Attribute-based Searchable Encryption (SC-ABSE) was proposed to solve the issue of security, privacy, and searchability in PHR [47]. This work bridges the gap between PHRs and blockchain technology by downloading extensive medical data into the IPFS and building a compulsory cryptography authorization and access control system for outsourced encrypted medical data. This system expands CP-ABE and searchable symmetric encryption (SSE), as well as using smart contract technologies, to accomplish the following: 1) efficient and secure fine-grained access control of outsourced encrypted data, 2) confidentiality of data by eliminating trusted private key generators, and 3) multi keyword searchable mechanism. The rigorous security indistinguishability analysis, based on decisional bilinear Diffie–Hellman hardness assumptions (DBDH) and dismulti-keywordhm (DL) issues, reveals that SC-ABSE is secure against the chosen-keyword attack (CKA) and keyword secrecy (KS) in the standard model. User collusion assaults are prevented, and data tamper-proof resistance is assured. Furthermore, security validation is validated by simulating a formal verification scenario with Automated Validation of Internet Security Protocols and

Applications (AVISPA), revealing that SC-ABSE is immune to man-in-the-middle (MIM) and replay attacks. Simulation findings demonstrate that SC-ABSE has high performance and low latency and that network life are ultimately increased in comparison with conventional medical systems.

Subsequently, the previous discussion of EHS with access control models aim to secure the system and protect privacy are summarized in the form of a comparative analysis. The explanation is shown in Table III.

Finally, all previous literature will be discussed in the form of an analytical discussion. The discussion is depicted in Table IV.

#### IV. SWOT ANALYSIS OF INSTRUMENTS CONSIDERABLE IN DEVELOPED COUNTRIES

In view of past work [1], an affordable EHS was developed and it was targeted to be utilized by the organizations in the developed countries. The instrument used to maintain and store EHR is by using a centralized database system. However, this storage is timely, frequent error, and costly [3], [4] to use and maintain, and it is against the aims of creating an affordable EHS in developed countries. As indicated in the preceding sections, blockchain system and cloud computing were the dominant instruments to be used for accessing, maintaining, and storing EHR. These instruments can also be used as one of the solutions for securing and protecting privacy. However, organizations in developed countries should have the knowledge to choose the right instruments for their new or upgraded EHS. Therefore, information about blockchain systems and cloud computing is provided in this paper in the form of strength, weakness, opportunity, and threat (SWOT) analysis to allow organizations doing instrument selection either to use only one or hybrid instruments. This analysis is essential planning that can take enormous amounts of information inside these four domains and sort out them into explicit concerns [48], [49]. Because of its successful and simple forms of analysis, hospitals may use this method, and it is ideal for use in strategic planning in healthcare systems or medical advances.

First, the Blockchain is discussed in the form of a SWOT analysis. The description of the analysis is discussed as follows and summarized in the form of a list in Table V.

##### 1) Strengths

Benefits were classified into two groups: patient-related benefits and organizational-related benefits.

For patient-related benefits, they include the followings:

- a) Users may only register their identity on the blockchain network once so they do not need to re-register their identification for the future [50].
- b) Allows healthcare professionals to embrace the concept of a shared database capable of producing sharable individualized healthcare plans for patients [50].
- c) The traceability feature enables tracking of the patient since every Bitcoin transaction is logged with a timestamp that is validated and

maintained by all computer nodes participating in the blockchain network [50].

- d) Enable effective patient monitoring, especially for critically sick patients, because this technology assists physicians in making appropriate medical-related treatment decisions. To do this, patients' wearable gadgets such as smartwatches, smartphones, and smart glasses must be linked to the healthcare blockchain network [50].
- e) Improve privacy protection for citizens and governments by giving individuals more control over their personal data. They can use blockchain technology to control who has access to their data, for what purpose, and for how long [51].

The organizational-related benefits are:

- a) To enable the secure sharing of patient information between healthcare organizations [50].
- b) To make clinical trial management easier because the study contains extremely sensitive patient-related information [50].
- c) The traceability function is crucial in controlling the pharmaceutical supply chain. In particular, can identify the origin of data, which can help pharmaceutical firms track the supply of products [50].
- d) Ability to manage medical insurance [50].
- e) Decentralized authority allows for the reduction of time, errors, and costs in the performance of processes, with the goal of building and updating a predictive model that supports medical care and risk management [3].
- f) The cryptographic system, the immutability of the data transmitted throughout the network, and the decentralized authority all contribute to increased confidence in the system [3].
- g) Every member can confirm the activities that happen in the organization as they have a duplicate of the entire blockchain on their gadget and this makes the process transparent [3].
- h) In Bitcoin, it is possible to identify any alteration to transaction records after they have been verified by solving a cryptographic problem [51].

##### 2) Weaknesses

- a) A verified transaction might be reversed after a government or group of persons in control of a blockchain [51].
- b) With the possibility for additional records to be produced natively on chain via smart contracts, legal admissibility must be considered. Laws and regulations regulating the admission and weight to be given to such evidence vary by jurisdiction, making it difficult to generalize how such evidence could be regarded by courts [51].

TABLE III. EHS WITH ACCESS CONTROL MODELS AIMS TO SECURE THE SYSTEM AND PROTECT PRIVACY

No.	Ref.	Environment		Mechanism										
		Cloud	Block Chain	MA- ABE	RBAC	MBAC	CP- ABE	ABSC	Cuckoo filter	PR- ABAC	ABE	SSE	AC	Channeling integrated
1.	Zhang, 2020	/		/										
2.	Azeez, 2018	/				/								
3.	Zarezadeh, 2020	/					/							
4.	Ming, 2018	/					/	/						
5.	Chenthara, 2019	/								/				
6.	Zhang, 2018	/					/							
7.	Riad, 2019	/												
8.	Tembhare, 2019	/			/						/			
9.	Nortey, 2019		/										/	/
10.	Dagher, 2018		/											
11.	Hussien, 2021		/				/					/		

- c) Blockchain records are intended to be immutable rather than changeable. This highlights the larger issue of how to remove or dispose of records from blockchain, which is also an issue to address when adopting data retention regulations or correcting mistakes in the record [51].
  - d) The topic of how to maintain blockchain records, in the long run, remains unanswered [51].
  - 3) Opportunities
    - a) As more companies see the advantages of blockchain-based recordkeeping, they will need a trusted adviser to assist them [51].
    - b) Through its usage, information professionals may reinvent their methods. For example, the UK National Archives is investigating smart contracts, to automatically execute data publishing [51].
  - 4) Threats
    - a) The sharing of whole copies of the blockchain under a model in which sensitive data on a single patient is shared would raise several issues related to privacy regulations, especially if entities other than public healthcare corporations participated in the network [52].
    - b) Scalability has become a major challenge for public blockchain applications, such as linking wearable devices because there is no control over the number of people joining the network [50].
    - c) Vulnerable to cyber-attacks where the attacker gains control of the blockchain network, which can lead to disaster if the attacker disrupts, stops or even reverses previously verified transactions within the network [50].
    - d) The high energy consumption has been noted since it pertains to the usage of the blockchain (public blockchain) based on proof of work, which is a mining process that uses a lot of electricity and it has gotten worse as more users have joined it and the number of transactions made per second has risen [50].
    - e) Absence of guidelines given by legitimate experts for blockchain advances [50].
    - f) Interoperability was considered as one of the significant difficulties for blockchain innovation reception in medical services because of the absence of trust between medical care organizations [50].
    - g) Lack of adequate technical skills and capabilities while carrying out blockchain advancements might prompt tragic results [50].
    - h) Data centers require a high cost of financing to maintain and require a large quantity of electricity [52].
    - i) One of the challenges is to train stakeholders on how to use this complex new system [52].
- Next, this paper describes the SWOT analysis of cloud computing. A detailed list of cloud analysis is indicated in Table VI.
- 1) Strengths
    - a) It has been recognized that cloud computing reduces the price of IT infrastructure, and the lower cost of IT infrastructure will open the road for certain businesses to embrace the technology [53].
    - b) Cloud computing allows companies to concentrate their efforts on their core competencies while also providing them with a scalability scenario, both in terms of services and infrastructure, that becomes “unlimited” [54].
    - c) Cloud resources can be anything: database services, virtual servers or machines, full service processes, or complex setups of distributed computing systems such as clusters [54].
    - d) It does not require hardware and software updates as it is managed by the cloud provider [55].
  - 2) Weaknesses
    - a) Concerns raised regarding the integrity, privacy, and security of services for users and their data [54].
    - b) Raises legal difficulties such as trademark infringement, security concerns, and the sharing of proprietary data resources [54].

- c) There is a lack of consistency in service legal agreement (SLA) terminology. Performance and availability are essential SLA goals, but additional variables like security, data (ownership, location, access, and portability), dispute mediation, disaster recovery, and exit strategy negotiation are also crucial [54].
- 3) Opportunities
  - a) Assist developed countries in reaping the advantages of cloud without the large upfront costs that have hindered previous attempts [54].
  - b) Many innovative services are produced in the cloud, such as educational applications for African or developed country schools [54].
  - c) Many vendors provide affordable cloud computing services [56].
  - d) Cloud computing research is still in its initial phases, particularly in the health industry [56].
  - e) The network, server, and security issues associated with locally installed, outdated systems are eliminated by adopting cloud computing [57].
- 4) Threats
  - a) Problem in terms of data security, IT audit policies [54].
  - b) Raises privacy problems since the service provider may access the data on the cloud at any moment, notwithstanding their own encryption claim. They may inadvertently or purposefully change or destroy data [54].
  - c) The problem is with the legal ownership of the data. Many Terms of Service agreements do not address ownership issues [54].
  - d) Lack of trust in cloud services [55].

## V. CONCLUSION

This paper discusses the EHS with access control to secure and preserve EHR. The issue addressed in this paper is about the EHS with the RBAC model. In general, RBAC is a prevalent model in access control, and it may still be used in current EHS research, despite the fact that it is considered a conventional access control model. The problem highlighted in this paper is that using EHS with RBAC only to secure and preserve an EHR may cause a huge risk to the system. Therefore, several of the current studies on EHS utilizing access control have been suggested and examined their mechanisms and environment for a notion to the organizations in developed countries to develop their EHS instead of using an autonomous role. Analytical discussion in the form of a table has also been provided to identify the issues or problems, findings or results, and comments or suggestions related to previous works. Finally, due to problems with the instrument used by the previous work, information on instrument selection was provided in the form of a SWOT analysis as it is hoped that this information can be useful for organizations in developed countries in obtaining ideas for building their new or upgraded EHS in the right environment.

In the future, further developments need to be considered. First, many different types of access control models were employed, such as trust, purpose, and attributes. Therefore, this is an opportunity for researchers to develop EHS with a variety of access control models instead of an anonymous role to secure the system and protect privacy. Second, instead of developing EHS in a blockchain and cloud environment, maybe developing EHS in another environment needs to be considered for example mobile or IoT environment.



No.	Ref.	Problems / Issues										Findings / Results	Comments / Suggestions		
		Data Access	Data Sharing	Patient Consent	Patient Control	Emergency Cases	Data Protection	Security Issues	CP-ABE Problems	Lack of RBAC	Other Problems				
16.	Lu, 2018													No experimental analysis	-
17.	Zhang, 2020		/		/									Comparison = Accomplishes privacy preservation	-
18.	Azeez, 2018		/											Show preliminary framework	Static
19.	Zarezadeh, 2020												/	Accomplishes data confidentiality, user anonymity, and collusion resistance.	-
20.	Ming, 2018		/											a) Security analysis = Secure b) Performance = Low cost, while maintaining the privacy	Hiding the AC policy may sacrifice efficiency.
21.	Chentthara, 2019		/											No experimental analysis	-
22.	Zhang, 2018											/		a) Security analysis = Secure b) Performance = Efficient and expressive than prior	Lack of revocation
23.	Riad, 2019											/		a) Security analysis = Secure and prevents unwanted access. b) Outstanding compatibility and performance.	Encryption schemes = Has a huge computational cost and inefficient for lightweight applications.
24.	Tembhare, 2019	/												Performance = Superior than prior	-
25.	Nortey, 2019											/		Transparency and privacy	No experimental analysis
26.	Dagher, 2018					/				/				No experimental analysis	High computational cost
27.	Hussien, 2021									/				a) Proved secured b) AVISPA = Immune to man-in-the-middle-attack and replay attack c) High performance	-

TABLE V. SWOT OF BLOCKCHAIN TECHNOLOGY

Strengths	Weaknesses
Patient-related: - Register once - Shared database - Traceability - Patient monitoring - Privacy protection  Organizational-related: - Secure sharing - Clinical trials - Pharmaceutical supply chain - Manage medical insurance - Reduce time, error, cost - Confidence - Transparent - Alteration detection	- Controlled by the top management - Legal admissibility - Disposition records - Maintain records
Opportunities	Threats
- Trusted advisor - Professional reinvention	- Sharing sensitive data - Scalability - Cyber-attack - High-energy consumption - Absence of guidelines - Inter-operability - Technical skills - Financial cost - Training

TABLE VI. SWOT OF CLOUD COMPUTING

Strengths	Weaknesses
- Reduce price - Scalability - Multi-purpose - Updated by the cloud provider	- Integrity, privacy, and security issues. - Legal issues - SLA inconsistency
Opportunities	Threats
- Assist developed countries - Produce innovative services - Services affordable - Research - Equipment and installation are removed	- Data security and audit - Privacy problem - Data legal ownership - Trust issues

REFERENCES

[1] O. E. Adetoyi and O. A. Raji, "Electronic health record design for inclusion in sub-Saharan Africa medical record informatics," *Scientific African*, vol. 7, p. e00304, 2020.

[2] W. Li, B. M. Liu, D. Liu, R. P. Liu, P. Wang, S. Luo, and W. Ni, "Unified fine-grained access control for personal health records in cloud computing," *IEEE Journal of Biomedical and Health Informatics*, vol. 23, no. 3, pp. 1278–1289, 2018.

[3] A. Fusco, G. Dicuonzo, V. Dell'Atti, and M. Tatullo, "Blockchain in healthcare: Insights on covid-19," *International Journal of Environmental Research and Public Health*, vol. 17, no. 19, p. 7167, 2020.

[4] T. Kanwal, A. A. Jabbar, A. Anjum, S. U. Malik, A. Khan, N. Ahmad, U. Manzoor, M. N. Shahzad, and M. A. Balubaid, "Privacy-aware relationship semantics-based XACML access control model for electronic health records in hybrid cloud," *International Journal of Distributed Sensor Networks*, vol. 15, no. 6, p. 1550147719846050, 2019.

[5] R. Sandhu, D. Ferraiolo, R. Kuhn *et al.*, "The NIST model for role-based access control: towards a unified standard," in *ACM Workshop on Role-based Access Control*, vol. 10, no. 344287.344301, 2000.

[6] H. A. Weber, "Role-based access control: the NIST solution," *SANS Institute InfoSec Reading Room*, 2003.

[7] S. Alshehri and R. K. Raj, "Secure access control for health information sharing systems," in *2013 IEEE International Conference on Healthcare Informatics*. IEEE, 2013, pp. 277–286.

[8] A. Small and D. Wainwright, "Privacy and security of electronic patient records—tailoring multimethodology to explore the socio-political problems associated with role based access control systems," *European Journal of Operational Research*, vol. 265, no. 1, pp. 344–360, 2018.

[9] H. Chi, E. L. Jones, and L. Zhao, "Implementation of a security access control model for inter-organizational healthcare information systems," in *2008 IEEE Asia-Pacific Services Computing Conference*. IEEE, 2008, pp. 692–696.

[10] P. Chinnasamy and P. Deepalakshmi, "HCAC-EHR: hybrid cryptographic access control for secure EHR retrieval in healthcare cloud," *Journal of Ambient Intelligence and Humanized Computing*, pp. 1–19, 2021.

[11] A. El Kettani, S. Housban, Z. Serhier, and M. B. Othmani, "Confidentiality in electronic health records systems: A review," *Journal of Medical and Surgical Research*, vol. 5, pp. 551–554, 2018.

[12] R. Gopalan, A. Antón, and J. Doyle, "UCONLEGAL: a usage control model for HIPAA," in *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium*, 2012, pp. 227–236.

[13] A. Lazouski, F. Martinelli, and P. Mori, "Usage control in computer security: A survey," *Computer Science Review*, vol. 4, no. 2, pp. 81–99, 2010.

[14] M. R. Salji, N. I. Udzir, M. I. H. Ninggal, N. F. M. Sani, and H. Ibrahim, "Trust, purpose, and role-based access control model for privacy protection," in *International Symposium on ICT Management and Administration (ISICTMA2019)*, 2019, p. 69.

[15] —, "Trust-based access control model with quantification method for protecting sensitive attributes," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 2, 2022. [Online]. Available: <http://dx.doi.org/10.14569/IJACSA.2022.0130282>

[16] S. Dixit, K. P. Joshi, and S. G. Choi, "Multi authority access control in a cloud EHR system with MA-ABE," in *2019 IEEE International Conference on Edge Computing (EDGE)*. IEEE, 2019, pp. 107–109.

[17] T. Tsegaye and S. Flowerday, "A Clark-Wilson and ANSI role-based access control model," *Information & Computer Security*, 2020.

[18] L. O. Nweke, P. Yeng, S. Wolthusen, and B. Yang, "Understanding attribute-based access control for modelling and analysing healthcare professionals' security practices," 2020.

[19] A. Tembhare, S. S. Chakkaravarthy, D. Sangeetha, V. Vaidehi, and M. V. Rathnam, "Role-based policy to maintain privacy of patient health records in cloud," *The Journal of Supercomputing*, vol. 75, no. 9, pp. 5866–5881, 2019.

[20] O. Alabi, "A review on information security of cloud based electronic health record," *Available at SSRN 3834180*, 2021.

[21] Y. Cheng, J. Park, and R. Sandhu, "Attribute-aware relationship-based access control for online social networks," in *IFIP Annual Conference on Data and Applications Security and Privacy*. Springer, 2014, pp. 292–306.

[22] W. Li, W. Ni, D. Liu, R. P. Liu, P. Wang, and S. Luo, "Fine-grained access control for personal health records in cloud computing," in *2017 IEEE 85th Vehicular Technology Conference (VTC Spring)*. IEEE, 2017, pp. 1–5.

[23] M. Price, P. Bellwood, N. Kitson, I. Davies, J. Weber, and F. Lau, "Conditions potentially sensitive to a personal health record (phr) intervention, a systematic review," *BMC medical informatics and decision making*, vol. 15, no. 1, pp. 1–12, 2015.

[24] J. Sun, X. Wang, S. Wang, and L. Ren, "A searchable personal health records framework with fine-grained access control in cloud-fog computing," *PLoS one*, vol. 13, no. 11, p. e0207543, 2018.

[25] S. Rana and D. Mishra, "Efficient and secure attribute based access control architecture for smart healthcare," *Journal of Medical Systems*, vol. 44, no. 5, pp. 1–11, 2020.

[26] Y. Zhang, M. Yang, D. Zheng, P. Lang, A. Wu, and C. Chen, "Efficient and secure big data storage system with leakage resilience in cloud computing," *Soft Computing*, vol. 22, no. 23, pp. 7763–7772, 2018.

[27] S. D. M. Satar, M. A. Mohamed, M. Hussin, Z. M. Hanapi, and S. D. M. Satar, "Cloud-based secure healthcare framework by using enhanced ciphertext policy attribute-based encryption scheme," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 6, 2021. [Online]. Available: <http://dx.doi.org/10.14569/IJACSA.2021.0120643>

[28] D. Suresh and M. L. Florence, "Securing personal health record system in cloud using user usage based encryption," *Journal of medical systems*, vol. 43, no. 6, pp. 1–11, 2019.

- [29] K. Edemacu, B. Jang, and J. W. Kim, "Efficient and expressive access control with revocation for privacy of PHR based on OBDD access structure," *IEEE Access*, vol. 8, pp. 18 546–18 557, 2020.
- [30] K. Seol, Y.-G. Kim, E. Lee, Y.-D. Seo, and D.-K. Baik, "Privacy-preserving attribute-based access control model for XML-based electronic health record system," *IEEE Access*, vol. 6, pp. 9114–9128, 2018.
- [31] A. Sathya and S. K. S. Raja, "Privacy preservation-based access control intelligence for cloud data storage in smart healthcare infrastructure," *Wireless Personal Communications*, vol. 118, no. 4, pp. 3595–3614, 2021.
- [32] A. R. Rajput, Q. Li, and M. T. Ahvanooy, "A blockchain-based secret-data sharing framework for personal health records in emergency condition," in *Healthcare*, vol. 9, no. 2. Multidisciplinary Digital Publishing Institute, 2021, p. 206.
- [33] T. T. Thwin and S. Vasupongayya, "Blockchain-based access control model to preserve privacy for personal health record systems," *Security and Communication Networks*, vol. 2019, 2019.
- [34] —, "Performance analysis of blockchain-based access control model for personal health record system with architectural modelling and simulation," *International Journal of Networked and Distributed Computing*, vol. 8, no. 3, pp. 139–151, 2020.
- [35] M. A. de Carvalho Junior and P. Bandiera-Paiva, "Strengthen electronic health records system (EHR-S) access-control to cope with GDPR explicit consent," *Journal of Medical Systems*, vol. 44, no. 10, pp. 1–7, 2020.
- [36] N. Huynh, M. Frappier, H. Pooda, A. Mammar, and R. Laleau, "SGAC : a multi-layered access control model with conflict resolution strategy," *The Computer Journal*, vol. 62, no. 12, pp. 1707–1733, 2019.
- [37] V. Rivera, "Formal verification of access control model for my health record system," in *2020 25th International Conference on Engineering of Complex Computer Systems (ICECCS)*. IEEE, 2020, pp. 21–30.
- [38] L. Zhang, Y. Ye, and Y. Mu, "Multiauthority access control with anonymous authentication for personal health record," *IEEE Internet of Things Journal*, vol. 8, no. 1, pp. 156–167, 2020.
- [39] N. A. Azeez and C. Van der Vyver, "Access control model for e-health in a cloud-based environment for HIV patients in South Africa," in *2018 IST-Africa Week Conference (IST-Africa)*. IEEE, 2018, pp. Page–1.
- [40] M. Zarezadeh, M. Ashouri-Talouki, and M. Siavashi, "Attribute-based access control for cloud-based electronic health record (EHR) systems," *The ISC International Journal of Information Security*, vol. 12, no. 2, pp. 129–140, 2020.
- [41] Y. Ming and T. Zhang, "Efficient privacy-preserving access control scheme in electronic health records system," *Sensors*, vol. 18, no. 10, p. 3520, 2018.
- [42] S. Chenthara, K. Ahmed, and F. Whittaker, "Privacy-preserving data sharing using multi-layer access control model in electronic health environment," *EAI Endorsed Transactions on Scalable Information Systems*, vol. 6, no. 22, 2019.
- [43] K. Riad, R. Hamza, and H. Yan, "Sensitive and energetic IoT access control for managing cloud electronic health records," *IEEE Access*, vol. 7, pp. 86 384–86 393, 2019.
- [44] R. N. Nortey, L. Yue, P. R. Agdedanu, and M. Adjeisah, "Privacy module for distributed electronic health records (EHRs) using the blockchain," in *2019 IEEE 4th International Conference on Big Data Analytics (ICBDA)*. IEEE, 2019, pp. 369–374.
- [45] G. G. Dagher, J. Mohler, M. Milojkovic, and P. B. Marella, "Ancile: Privacy-preserving framework for access control and interoperability of electronic health records using blockchain technology," *Sustainable cities and society*, vol. 39, pp. 283–297, 2018.
- [46] Y. Lu, R. O. Sinnott, K. Verspoor, and U. Parampalli, "Privacy-preserving access control in electronic health record linkage," in *2018 17th IEEE International Conference on Trust, Security and Privacy in Computing and Communications/12th IEEE International Conference on Big Data Science and Engineering (TrustCom/BigDataSE)*. IEEE, 2018, pp. 1079–1090.
- [47] H. M. Hussien, S. M. Yasin, N. I. Udzir, and M. I. H. Ninggal, "Blockchain-based access control scheme for secure shared personal health records over decentralised storage," *Sensors*, vol. 21, no. 7, p. 2462, 2021.
- [48] B. Phadermrod, R. M. Crowder, and G. B. Wills, "Importance-performance analysis based SWOT analysis," *International Journal of Information Management*, vol. 44, pp. 194–203, 2019.
- [49] S. Walston, *Strategic Healthcare Management: Planning and Execution, Second Edition*, ser. AUPHA/HAP Book. Health Administration Press, 2018. [Online]. Available: <https://books.google.com.my/books?id=DocxQEACAAJ>
- [50] I. Abu-Elezz, A. Hassan, A. Nazeemudeen, M. Househ, and A. Abd-Alrazaq, "The benefits and threats of blockchain technology in healthcare: A scoping review," *International Journal of Medical Informatics*, p. 104246, 2020.
- [51] V. L. Lemieux, "Blockchain recordkeeping: A SWOT analysis," *Information Management*, vol. 51, no. 6, pp. 20–27, 2017.
- [52] S. Alla, L. Soltanisehat, U. Tatar, and O. Keskin, "Blockchain technology in electronic healthcare systems," in *Proceedings of the 2018 IISE Annual Conference*, 2018, pp. 1–6.
- [53] O. M. S. H. Ali and A. Shrestha, "Analysis of the total cost of ownership for cloud computing technology adoption: A case study of regional municipal government sector," no. 56, 2017. [Online]. Available: <https://aisel.aisnet.org/acis2017/56>
- [54] M. M. Seke, "Be mindful of the move: A SWOT analysis of cloud computing towards the democratization of technology," *i-manager's Journal on Cloud Computing*, vol. 5, no. 1, p. 26, 2018.
- [55] J. Singh, "Study on challenges, opportunities and predictions in cloud computing," *International Journal of Modern Education and Computer Science*, vol. 9, no. 3, p. 17, 2017.
- [56] S. J. Putra, M. N. Gunawan, D. P. Sari, S. Ratnawati, Y. Sugiarti *et al.*, "A cloud computing based for clinical information system," 2020.
- [57] M.-H. Kuo, A. Kushniruk, and E. Borycki, "Can cloud computing benefit health services? : A SWOT analysis," in *User Centred Networked Health Care*. IOS Press, 2011, pp. 379–383.

# Semi-supervised Method to Detect Fraudulent Transactions and Identify Fraud Types while Minimizing Mounting Costs

Chergui Hamza<sup>1</sup>, Abrouk Lylia<sup>2</sup>, Cullot Nadine<sup>3</sup>, Cabioch Nicolas<sup>4</sup>  
Université de Bourgogne, SKAIZen Group, 4 av. Alain Savary,  
21000 Dijon<sup>1,2,3</sup>  
SKAIZen Group, 14 rue de mantes,  
92700 Colombes<sup>4</sup>

**Abstract**—Financial fraud is a complex problem faced by financial institutions, and existing fraud detection systems are often insufficient, resulting in significant financial losses. Researchers have proposed various machine learning-based techniques to enhance the performance of these systems. In this work, we present a semi-supervised approach to detect fraudulent transactions. First, we extract and select features, followed by the training of a binary classification model. Secondly, we apply a clustering algorithm to the fraudulent transactions and use the binary classification model with the SHAP framework to analyze the clusters and associate them with a particular fraud type. Finally, we present an algorithm to detect and assign a fraud type by leveraging a multi-fraud classification model. To minimize the mounting cost of the model, we propose an algorithm to choose an optimal threshold that can detect fraudulent transactions. We work with experts to adapt a risk cost matrix to estimate the mounting cost of the model. This risk cost matrix takes into account the cost of missing fraudulent transactions and the cost of incorrectly flagging a legitimate transaction as fraudulent. In our experiments on a real dataset, our approach achieved high accuracy in detecting fraudulent transactions, with the added benefit of identifying the fraud type, which can help financial institutions better understand and combat fraudulent activities. Overall, our approach offers a comprehensive and efficient solution to financial fraud detection, and our results demonstrate its effectiveness in reducing financial losses for financial institutions.

**Keywords**—Machine learning; semi-supervised learning; fraud; finance; cost analysis

## I. INTRODUCTION

Financial institutions face multiple challenges in fighting money laundering activities. Jensen [1] defines it as *the disguise of the origin of illegally obtained funds to make them appear legitimate. The goal of money laundering is to convert cash into another form.* Financial institutions must fight fraudulent activities by analyzing their customers' transactions. Transactions exchanged between financial institutions use SWIFT (Society for Worldwide Interbank Financial Telecommunication). This provides an interbank network offering different services, such as money transfers between bank accounts. More than 11,000 banking organizations across nearly 200 countries use SWIFT to transfer money [2]. SWIFT transactions are international and may have multiple intermediaries between the transaction's originator and beneficiary. However, among these transactions, anomalies may be linked

to financial fraud. Thus, the analysis of interbank transactions is a crucial issue for financial institutions [3]. The current systems have four limitations outlined by SWIFT<sup>1</sup>: 1) **systems and processes inefficiency**, originally designed for retail banking, is based on rules and risks. In this context, a critical problem is the high alert volume generated by them, and 90 percent of them are false alerts. Therefore, a manual investigation by experts is required. 2) Due to **mounting costs**, financial institutions lose a considerable amount of money fighting against money laundering—either by being fined for their weak compliance systems, maintaining these, or paying experts to review the alerts. 3) **Indirect structure**: Some domestic and regional banks act as aggregators for smaller banks. It is hard to follow and monitor the payment activity effectively in such instances. 4) **Information sharing**: Banks do not share information with their customers due to confidentiality clauses. In addition, due to these limitations, fraudulent transactions' detection is a complex task. It's also difficult to identify the fraud types [4]. For these reasons, our work aims to respond to the first three challenges by answering the following questions: How can fraudulent transactions be detected with their frauds types while minimizing the mounting costs? The article is structured as follows: in Section II, we review the most recent developments in the field of machine learning for detecting financial fraud and identifying different types of fraud. Then, in Section III, we present our method, which aims to detect fraudulent transactions within the SWIFT network and identify fraud types, while minimizing the associated costs. In Section IV, we apply our methodology to a real-world dataset and report on the experimental results, demonstrating the effectiveness of our approach. Finally, in Section VI, we summarize our findings and outline potential avenues for future research.

## II. RELATED WORK

Most of the detection systems are based on rules. These systems, based on predefined rules pertaining to amounts, countries, or customer behaviors, are identified by fraudsters. In turn, they adapt their behaviors to bypass these rules and manage to launder their money through illegal activities. For these reasons, they generate a high false alerts rate and detect few fraudulent transactions. Most of the literature on

<sup>1</sup><https://www.swift.com/fr/node/166756>

financial fraud detection techniques are based on machine learning. These techniques “offer numerical power and functional flexibility needed to identify complex patterns” [5]. Machine learning techniques for detecting financial fraud follow a process that can be divided into four steps: 1) **data acquisition**, which can be a complicated task for specific fields, such as finance or medicine, in instances where the data is confidential. Synthetic data can be used to validate experiments. 2) **Features extraction** involves calculating new information from existing features and reducing the number of dimensions while storing the information in the initial features. The 3) choice of **algorithms and their hyperparameters** depends on the number of dimensions, the volume, and the nature of the data. 4) The last step is the model **evaluation**. The predictive models are evaluated with metrics based on correct and wrong predictions.

This section is structured as follows: we present the machine learning process used for fraudulent transaction detection. Then we present a fraud types identification work. Finally, we synthesize the related work and introduce our approach.

#### A. Machine Learning Process

1) *Data acquisition*: A financial dataset is transactions with fields such as the amount, the date, the beneficiary [6], [7], [8]. They are confidential, and the lack of public data hinders experiments, particularly for their validation and comparison. There are different data formats depending on the data source (retail bank [9], agricultural bank [10]). The data volume in various experiments is heterogeneous, ranging from thousands [11] to millions [12] transactions. Validation of fraud detection approaches requires labeled datasets with legitimate and fraudulent transactions. A common aspect of financial fraud datasets is the class imbalance between fraudulent and legitimate transactions, with a fraudulent ratio usually around 0.1% [13]. In this context, researchers are interested on synthetic financial data generation. Lopez and al. [14] developed *PaySim* a mobile payment simulator tool with fraudulent transactions.

Most of the studied approaches don't explain the data generation process [15], [10]. Michalak and al. [16] detail in their study how they generated transactions. They use Gaussian distribution to generate a transactions networks between employees of companies. LV and al. [17] use real data coupled with artificially generated fraudulent data. Some approaches use the *kaggle* public dataset<sup>2</sup> to validate anti-money laundering methods [18], [7], [19]. This dataset comprises transactions conducted with credit cards involved in fraud. It includes the following three known attributes: the date, the amount, the transaction class (fraudulent or legitimate), and 28 remaining attributes with unknown meanings.

2) *Features extraction*: The data must be processed before training the models with algorithms. There are many types of processing, namely standardization, features addition, or dimensions reduction. Features from the transaction attributes are extracted to represent customers' behaviors based on their transaction history. The transactions are aggregated with different periods (weekly, monthly, and yearly) to compute several features, such as the transactions' average amounts made by customers and their frequency (number of transactions

done in a period). There are few works based on SWIFT transactions [20] or international transactions fields (countries or currencies) [21] to extract features. When the features number is too high, a dimensions reduction step is used to train the models faster or for visualization purposes. Bestami et al. [19] use the PCA (Principal Component Analysis) algorithm to reduce dimensions number to train a model with the K nearest neighbors algorithm. Paula et al. [22] use auto-encoders (deep learning technique) to reduce dataset dimensions number and complete the training 20 times faster. As mentioned, financial fraud datasets are imbalanced. The class imbalance problem can be less constraining using over or under-sampling techniques. Oversampling techniques generate synthetic new instances from the minority class, whereas under-sampling is used to reduce the number of instances from the majority class. SMOTE [23] is a popular oversampling algorithm in the literature that generate additional fraudulent transactions from the existing ones in the dataset. Badal et al. [24] prove this technique effective in financial fraud detection by obtaining better results using the SMOTE algorithm.

3) *Algorithms and hyperparameters*: Fraud analysts use fraud detection rules. These rules are used to detect fraudulent scenarios that occur frequently. However, rules can become quickly obsolete and must be reviewed. Nowadays, machine learning can be combined with a rules-based system. Classification (supervised learning) and clustering (unsupervised learning) help in fraud prevention and detection by classifying transactions as fraudulent or legitimate. The choice between supervised and unsupervised learning depends on the datasets. Supervised learning aims to learn the relationship between the data and its label. In unsupervised learning, the goal is to retrieve exploratory information, by grouping similar data or detecting hidden patterns [5]. Ryman-Tubb et al. [25] conduct a survey on card fraud detection methods for using financial transactions. This survey shows that only eight methods can be deployed on real data. Al-Hashedi et al. [26] expanded the work of Albashrawi [27] and present a survey from 2009 to 2019 conducted on financial fraud classified by fraud types.

a) *Supervised methods*: These compare algorithms to deduce which is pertinent to the data and their volume [28], [29], [19]. Mehbodniya et al. [30] propose financial fraud detection in healthcare based on machine learning and deep learning techniques and showed that the KNN algorithm generates better results than other approaches. Ensemble methods such as random forest [31] or boosting algorithms [32], [33], which combine multiple models, also proved their effectiveness in imbalanced datasets by using local decisions taken in areas where the imbalance is less prominent.

b) *Unsupervised methods*: These are used for financial fraud detection. Porwal et al. [7] use K-means algorithm to create fraudulent and legitimate transaction clusters. Simultaneously, other works propose new distance measurements to detect outliers [34]. Guo et al. [35] use autoencoders to have a deep representation of their data; they combined it with a KNN-based outlier detection method [36].

c) *Semi-supervised methods*: This combines supervised and unsupervised learning. Some approaches [37], [38], [8], [39] apply unsupervised algorithms to label their data, then they use supervised algorithms to classify their transactions. These approaches don't need an expert to verify each abnormal

<sup>2</sup><https://www.kaggle.com/mlg-ulb/creditcardfraud>

transaction from the unsupervised model, they verify clusters and fraudulent transactions from the supervised model.

4) *Evaluation and metrics:* The trained models are evaluated to check their effectiveness. Many metrics for evaluation, such as precision, recall, or F1-Score, exist. These metrics are used for model validation and comparison. In the evaluation process, data volume and fraudulent class rate are important. The evaluation should be done on the minority class.

In the case of unlabeled datasets, the evaluation is more complex. The verification of prediction results can be a long and imprecise operation. Furthermore, in the financial domain, Bahnsen et al. [40] propose a cost-risk matrix (Table I) to estimate the model mounting cost for a financial institution fighting against credit card fraud. It estimates model mounting costs depending on its predictions: it has an administration cost  $C_a$  for transactions predicted fraudulent, representing the estimated price for a transaction investigated by an expert.  $Amt_i$  is a fraudulent transaction cost, the fraudulent transaction amount predicted as legitimate by the model. They sum up all the transaction costs used in the evaluation phase to compute the model mounting costs.

TABLE I. BAHNSEN [40] RISK-COST MATRIX

		Reality ( $t_i$ )	
		Fraud	Legit
Prediction ( $t_i$ )	Fraud	$C_a$	$C_a$
	Legit	$Amt_i$	0

### B. Fraud Types Identification

Desrousseau et al. [41] present an approach to profile money laundering activities. They use the SOM (Self-Organizing Map) algorithm from an unlabeled transaction dataset to map its transactions into a two-dimensions matrix. SOM algorithm [42] is used as an unsupervised algorithm to cluster and visualize data with a two-dimensional map. The algorithm assign a new representation for each transaction. It uses the map node value associated to the transaction. Desrousseau et al. use this node and its neighborhood as a new transaction representation. With it, they train a neural network called Fuzzy ART, which forms clusters with transactions. Finally, they combine these clusters with the map from the SOM algorithm, resulting in a map with different regions depending on the value of the node associated with a cluster. To interpret results, they use two methods: 1) They use the Fuzzy Art model weighted vector for each money laundering type and features distribution to retrieve the features with the highest weight. 2) They group transactions with the same type and use features distribution. This approach is interesting because no literature interprets fraud types in financial datasets. The choice of algorithms might be questionable because there are numerous clustering algorithms, such as k-means or BIRCH. The interpretation through the comparison of maps can be very complicated depending on the number of dataset features. Moreover, this approach relies on unlabeled datasets. It does not address the banks' concern of identifying the fraudulent patterns on their labeled dataset and the fraud types identification. While their interpretation of feature distribution is interesting, a new tool in the literature has been designed to interpret models, which could be helpful. In our work, we aim

TABLE II. EXAMPLE OF SWIFT MESSAGES OF THE DATASET

Originator	Intermediary	Beneficiary	Date	Currency	Amount	Class
BIC0FR01	BIC0IT01	BIC0FR02	210625	EUR	15006	L
BIC0US03	-	BIC0GB01	210625	GBP	33065	L
BIC0FR04	BIC0FR06	BIC0FR05	210626	EUR	100325	F

to extend this work, propose a method to detect fraudulent transactions with a reduced features number, and interpret the fraud types with interpretable tools.

### C. Synthesis

We studied the related works of fraudulent transactions detection and types identification. This problem has not been considered yet on the SWIFT transactions. As mentioned, SWIFT executes financial transactions between banks worldwide. These transactions have fields such as: country, currency and intermediary. SWIFT fraudulent transactions detection model is obtained through a comparative study of the supervised and unsupervised algorithms and their evaluation.

Desrousseau et al. [41] present interesting results to identify the fraud types inside our datasets. However, this approach could be improved by considering labeled datasets and using model interpretation (SHAP). We aim to reduce our model mounting costs by using and expanding the risk-cost matrix from [40].

## III. METHODOLOGY

In this section, we present our proposed approach for labeled transactions (fraudulent and legitimate). Our approach has three goals:

- 1) Detect fraudulent transactions
- 2) Identify and analyze the fraud types in the dataset
- 3) Minimize the mounting costs

To achieve these goals, we present our semi-supervised approach in Fig. 1. First, we extract features from a SWIFT transaction set. Second, we select the relevant features for detecting fraudulent transactions from the legitimate by training a binary classification model. Afterward, we create a fraudulent transaction subset, from which we apply an unsupervised algorithm to form clusters based on the fraud types. We also use the binary classification model to interpret the cluster to identify the fraud type. Moreover, we label fraudulent transactions based on the fraud types and train a multi-class classification model. We propose the *Multi-Fraud Detection* algorithm (MFD) to classify a transaction as legitimate or classes associated to fraud types. Then we propose a second algorithm to minimize the model mounting cost.

This section is structured as follows: (A) we present the dataset and the fields, (B) we list the extracted features, (C) we train a binary classification model and evaluate it, (D) we apply a clustering algorithm on the fraudulent transactions and identify the fraud type, (E) we train a multi-class classification model, and (F) we minimize the model mounting costs.

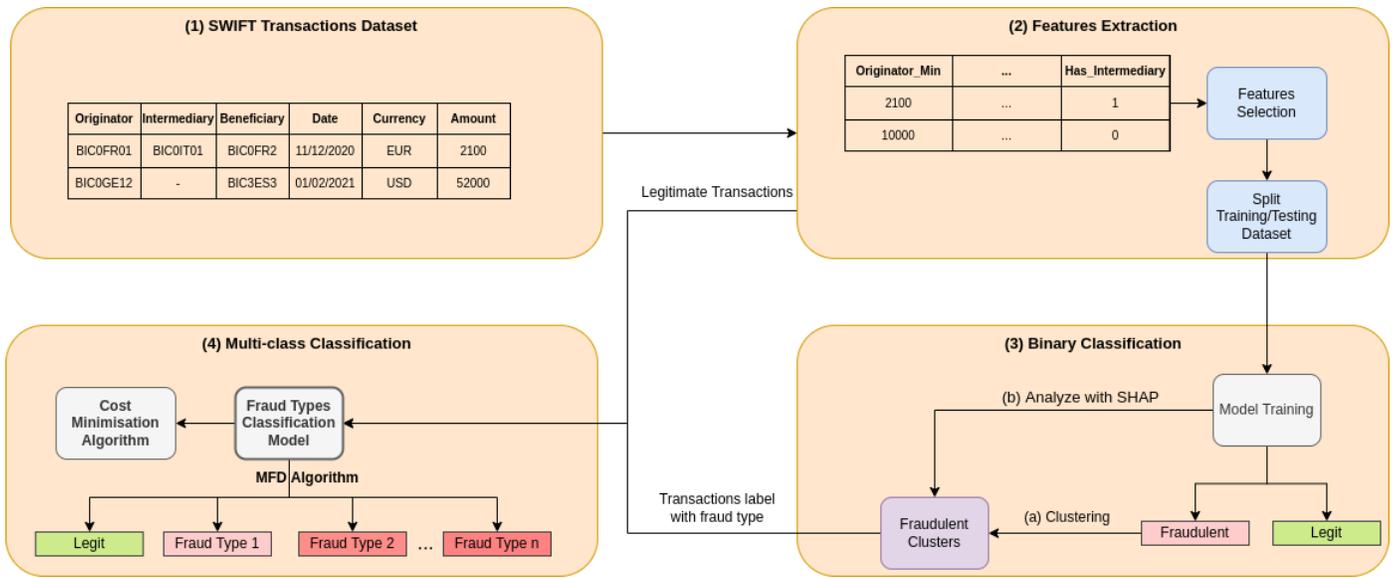


Fig. 1. Methodology architecture schema

### A. Dataset Presentation

In Table II, we present the SWIFT transactions fields. These transactions contain three actors: the originator, the intermediary, and the beneficiary. The transaction corresponds to the transfer of a money **amount** in a **currency** operated on a **date** between an **originator** and a **beneficiary**. SWIFT transactions path depends on the relationship between the **originator** bank and the **beneficiary** bank. If they have no direct relationship, then the transaction involves an **intermediary** connecting the two banks. Otherwise, the path only includes the originator and the beneficiary. An actor is identified with a BIC code, which contains the financial institution country. The transaction class indicates if the transaction is fraudulent (F) or legitimate (L). We formalize transactions as the following:

We have a set  $T$  of transactions  $t_i$ , where  $i$  ranges from 1 to  $n$ , the total number of transactions. To compute features for each transaction  $t_i$ , we form transactions subset with a similar field (e.g. originator), with specific time window, (e.g. last 15 days). For example, for a transaction  $t_1$ , we have subset  $T^{15D}(\text{originator})$  containing all transactions with the same originator for the last 15 days.

### B. Detection of Fraudulent Transactions

We need to extract features from the dataset to separate legitimate and fraudulent transactions. The computed features must enlighten the fraud associated with SWIFT transactions. In a transaction, we have one numerical field: the amount. Features are computed on different period with the date field. Hence, we use the amount and the date to compute features associated with the other fields: the originator, the intermediary, the beneficiary and their countries, and the currency. With a financial expert, we define a list of features with their formalization in Table III.

We also define other features relative to the transaction path. In this context, by path, we mean the countries implied

in the transaction and the presence of an intermediary. We list the features in Table IV.

Finally, we add temporal features relative to the day, the week, the year. We select a reduced features number. Some features might be irrelevant to the separate legitimate and fraudulent transactions in the features computed. A feature selection algorithm assigns an importance value to each feature. We select the top  $n$  features depending on the feature number required to interpret the fraud type. With the additional features, we train a binary classification model, evaluated with the metrics presented in Section III-C.

### C. Model Training and Evaluation

In the related works, researchers trained a model with different classifiers and then selected the classifier with the best evaluation results. A good evaluation is crucial to ensure that the extracted features can distinguish between legitimate and fraudulent transactions. To do this, we use a confusion matrix presented in Table V.  $TP$  is the number of true positives.  $FN$  is the number of false negatives.  $FP$  is the number of false positives. Finally,  $TN$  is the number of true negatives.

From the confusion matrix, we can evaluate our approach with the classical metrics:

$$Precision = \frac{TP}{TP + FP}; \quad (1)$$

$$Recall = \frac{TP}{TP + FN}; \quad (2)$$

$$F1 = \frac{2 * TP}{2 * TP + FP + FN}; \quad (3)$$

### D. Clustering of Fraudulent Transaction

Unsupervised algorithms are effective at discovering new patterns and hidden relations in datasets. Thus, we create

TABLE III. FIELDS BASED FEATURES

Formalization	Description
$ T(field) $	Transactions number in the set
$\max_{0 < i < n} T(field amount)$	Maximum amount
$\min_{0 < i < n} T(field amount)$	Minimum amount
$\sum_{i=0}^t T(field amount)$	Sum of transactions
$\frac{\sum_{i=0}^t T(field amount)}{ T(field) }$	Average transactions amounts
Latency	Seconds since the last transaction
$ T(field) \cap T(Originator) $	Count of transactions with the same Originator
$ T(field) \cap T(Intermediary) $	Count of transactions with the same Intermediary
$ T(field) \cap T(Beneficiary) $	Count of transactions with the same Beneficiary
$ T(field) \cap T(Currency) $	Count of transactions with the same Currency
$ T(field) \cap T(OriginatorCountry) $	Count of transactions with the same Originator Country
$ T(field) \cap T(IntermediaryCountry) $	Count of transactions with the same Intermediary Country
$ T(field) \cap T(BeneficiaryCountry) $	Count of transactions with the same Beneficiary Country
$ \{T(field) \cap T(Originator)\} $	Distinct count of transactions with the same Originator
$ \{T(field) \cap T(Intermediary)\} $	Distinct count of transactions with the same Intermediary
$ \{T(field) \cap T(Beneficiary)\} $	Distinct count of transactions with the same Beneficiary
$ \{T(field) \cap T(Currency)\} $	Distinct count of transactions with the same Currency
$ \{T(field) \cap T(OriginatorCountry)\} $	Distinct count of transactions with the same Originator Country
$ \{T(field) \cap T(IntermediaryCountry)\} $	Distinct count of transactions with the same Intermediary Country
$ \{T(field) \cap T(BeneficiaryCountry)\} $	Distinct count of transactions with the same Beneficiary Distinct country

TABLE IV. PATH FEATURES

Features	Description
Intermediary	Boolean value to specify if there is an intermediary
Originator path	The count of the originator using the intermediary and beneficiary countries to make a transaction
Intermediary path	The count of the originator using the originator and beneficiary countries to make a transaction
Beneficiary path	The count of the originator using the originator and intermediary countries to make a transaction
Distinct originator Path	The count of a distinct path with an intermediary and beneficiary countries
Distinct intermediary Path	The count of a distinct path with an originator and a beneficiary countries
Distinct beneficiary path	The count of a distinct path with an originator and an intermediary countries

TABLE V. CONFUSION MATRIX

	Predicted : Positive	Predicted : Negative
Actual : Positive	$TP$	$FN$
Actual : Negative	$FP$	$TN$

fraudulent transactions clusters, each of which will be associated with a fraud type. Cluster analysis allows experts to identify which fraud types clusters are associated based on their fraud knowledge. We use the SHAP framework [43] to facilitate cluster analysis and interpret model predictions. It assigns importance to each feature (Shapley values) for a prediction, depending on the feature’s value. By leveraging this framework, we use visualization tools: *heatmap* and *beeswarm* for each cluster.

### E. Multi-Class Model

After identifying the fraud types, we train a new model to predict the transactions class between legitimate and fraud types. Class numbers rely on the fraud types’ numbers on the dataset. The new model attributes a probability to each

class; however, the frauds’ probability is split into different classes. Some fraudulent transactions’ probability is divided into different fraudulent classes. The legitimate class could, in turn, take over them. For these reasons, we sum the fraudulent classes’ probability  $p_i$  ( $1 < i < n$ ,  $n$  number of fraudulent classes), we start at 1 because  $i = 0$  represents the legitimate class. The transaction is considered fraudulent if the sum of  $p_i$  is above a defined threshold. The transaction’s class will be the fraudulent with the highest probability. We resumed this on our *MFD* algorithm presented in the Algorithm 1.

### F. Mounting Costs Minimization

To estimate the model mounting costs, we used the matrix of Bahnsen et al. [40], where we added a cost  $C_d$  for the legitimate transaction predicted as fraudulent, which estimates the dissatisfaction price of a customer whose transaction has been blocked. Indeed, Bahnsen et al. did not consider the wrong prediction cost for a fraudulent transaction. If a transaction is incorrectly blocked for a customer, then this one could be unsatisfied and result in losses for the financial institution.

We minimize the model’s risk cost with our new risk–cost

**Algorithm 1** MFD : Multi-Fraud Detection

```

1: T : set of transactions
2: m : classification model
3: threshold : fraudulent threshold
4: n : fraudulent classes number
5: for t in T do
6:   p = model.predict_proba(t)           ▷ list of pi
7:   sum = ∑i=1n pi
8:   if sum > threshold then
9:     fraudtype_index = argmax(p)
10:    class_list.add(fraud[fraudtype_index])
11:   else
12:     class_list.add(legit)
13:   end if
14: end for
15: return class_list

```

TABLE VI. ADAPTED RISK-COST MATRIX

		Reality (t <sub>i</sub> )	
		Fraud	Legit
Prediction (t <sub>i</sub> )	Fraud	C <sub>a</sub> + C <sub>d</sub>	C <sub>a</sub>
	Legit	Amt <sub>i</sub>	0

matrix in Table VI and Algorithm 2. For that we use the *MFD* algorithm for each threshold value between 0 and 1 with a 0.01 step. Subsequently, with the model's prediction with these thresholds, we compute the model, then, we add f1-score and cost in list. Afterward, we retrieve the highest f1-score and check the closest f1-score with the lowest cost. Moreover, we return the threshold corresponding to this, and by doing so, we ensure to keep our model effective while reducing its mounting costs.

**Algorithm 2** Cost Minimization

```

1: T : set of transactions
2: ca : administrative cost
3: cd : dissatisfaction costs
4: amt : transactions' amounts list
5: class_target : transactions' class
6: model : trained model
7: f1_list : f1-score list
8: costs_list : costs list
9: for threshold ← 0 to 1 by 0.01 do
10:   class_list = MFD (T, model, threshold)
11:   f1_list.add(compute_f1(class_list,class_target))
12:   costs_list.add(compute_costs(class_list,
    class_target,ca,cd,amt))
13: end for
14: best_f1 = max(f1_list)
15: thresholds_around_best_f1 = around(best_f1, f1_list)
16: best_threshold = argmin_cost(costs_list, thresholds_around_best_f1)
17: return best_threshold

```

IV. EXPERIMENTATION

Experiments were conducted with a 3676795 SWIFT transactions dataset obtained through a collaboration with the

TABLE VII. BASE FIELDS NAME

Originator
Intermediary
Beneficiary
Common history between Originator and Intermediary
Common history between Originator and Beneficiary
Common history between Intermediary and Intermediary
Currency
Originator country
Intermediary country
Beneficiary country

TABLE VIII. THE 10 FEATURES SELECTED

Features Name
Value
number with intermediary Beneficiary 3D
max value Originator Beneficiary
avg value Originator Beneficiary
avg value Intermediary Beneficiary
latency Intermediary Beneficiary
frequency with currency Intermediary Beneficiary
sum value Intermediary Beneficiary 3D
max value Intermediary Beneficiary 3D
frequency with currency Intermediary Beneficiary 3D

SKAIZen Group<sup>3</sup> company. We split the data into a training dataset composed of 294136 transactions with 10722 fraudulent transactions and a testing dataset composed of 735359 transactions with 2645 fraudulent transactions. Thereafter, we used the Jupyter<sup>4</sup> platform to develop the experimentation. We trained our models with the Scikit-Learn library.

A. Features Extraction

We compute the features for the fields listed in Table VII.

Then, we compute the features related to transactions paths. Features are extracted for each transaction on a time window of one year, one month, and three days before the transaction date. We obtain 267 features, and we reduce this number using the *SelectFromModel* algorithm, a meta-transformer with a model trained with a classifier to assign an importance score to each feature. We use CatBoost as the classifier for its performance and capacity to deal with categorical features. The features number is an algorithm parameter, we choose 10 features based on SWIFT experts' knowledge (Table VIII).

The algorithm selects features related to the relationship between the intermediary and the beneficiary. For the transactions' history, no features related to the month were retained, except the last three days ('3D' suffix) and the last year. We apply a value transformation between 0 and 1 with the *Quantile Transformer* algorithm, which transforms according to a uniform distribution to reduce the outlier transaction impact. A very high amount of transactions could misrepresent other transactions during the visualization step (section IV-C).

B. Algorithms Comparison

We compare the classifiers from the literature to observe the best algorithm for our data. The results are presented in Table IX. Results show that CatBoost is the best algorithm for our data with a f1-score of 0.89.

<sup>3</sup><https://skaizengroup.eu/>

<sup>4</sup><https://jupyter.org/>

TABLE IX. CLASSIFIERS COMPARATIVE ACCORDING TO PRECISION, RECALL AND F1-SCORE

	Fraudulent			Legitimate			All(average)		
	Precision	Recall	F1-Score	Precision	Recall	F1-Score	Precision	Recall	F1-Score
SVM	1.0	0.08	0.16	0.99	0.99	0.99	0.99	0.54	0.58
Random Forest	0.93	0.61	0.74	0.99	0.99	0.99	0.97	0.80	0.87
LightGBM	0.70	0.55	0.62	0.99	0.99	0.99	0.85	0.77	0.81
XGBoost	0.92	0.61	0.74	0.99	0.99	0.99	0.96	0.80	0.87
CatBoost	0.94	0.68	0.79	0.99	0.99	0.99	0.96	0.84	<b>0.89</b>

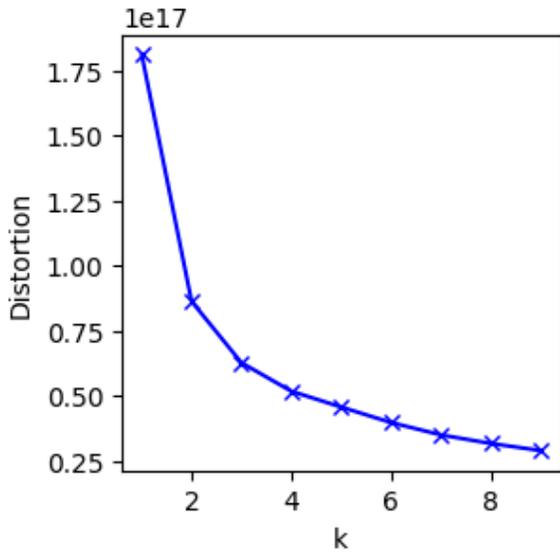


Fig. 2. Elbow Figure

TABLE X. CLUSTERS DISTRIBUTION

Cluster	Count of transactions	Average
0	4970	107999
1	4652	771069
2	3745	1279286

Our features are relevant enough to distinguish legitimate and fraudulent transactions. Our future experiment models will be trained with CatBoost.

C. Clustering of Fraudulent Transactions

After algorithms comparison, k-means [44] unsupervised algorithm is applied on the fraudulent transactions subset. We test different cluster numbers on our 13367 fraudulent transactions. We used the elbow method [45] to select the optimal  $k$  Fig. 2. The optimal cluster number is 3 because the curve linearly decreases at this number. Table X presents transactions distribution for each cluster and the average of transactions amount. The 3 clusters have an equivalent transactions number; however, the average of the transactions amount is different. The first cluster has a low average, the second one has a medium average, and the third has a high average.

In order to assign a fraud type to each cluster, we present fraud types in Table XI identified by our experts and the literature [46].

TABLE XI. FRAUD TYPES

Fraud Types	Description
Payment diversion	unauthorized redirection of payment instructions
Large amount	unauthorized initiation of high-value transactions
Smurf	illegal money laundering using money mules
Dormant account	unauthorized use of inactive bank accounts
False demand draft	creation and use of fake financial instruments to withdraw money.

D. Clusters Analyses and Fraud Types Identification

To analyze clusters, we used the SHAP framework [43] with the binary classification model trained on section IV-B on the transactions of each cluster. We compute the Shapley values with SHAP. For the interpretation, we use two visualization types: i) *beeswarm* for global cluster visualizations with features' importance and their value, ii) *heatmap* for cluster local visualizations with features impact on each transaction prediction.

1) *Cluster 0*: From the *heatmap* (Fig. 3), we split the map in two parts with the first five features : (i) on the left side, the first three features have a high impact. The *beeswarm* (Fig. 4) indicates that these values are low. For the last two features, their impact is lower, and their values are medium. (ii) The right side indicates the high impact of the first and fourth features with low values and a negative impact of the feature *value* (Amount field) with low values. This fraud type comprises customers making a few transactions with low and medium amounts.

Experts analyzed it as either to new customers doing few transactions (low frequency) of low amount (left side), or either to customers doing medium amount transaction after a long time (right side). This cluster is assigned to the “dormant account” fraud type (DAF).

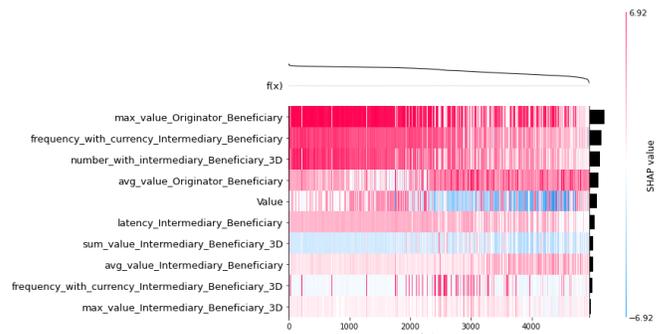


Fig. 3. Heatmap cluster 0

2) *Cluster 1*: According to the *heatmap* (Fig. 5), we split the map in two parts: (i) the right side shows that the amount

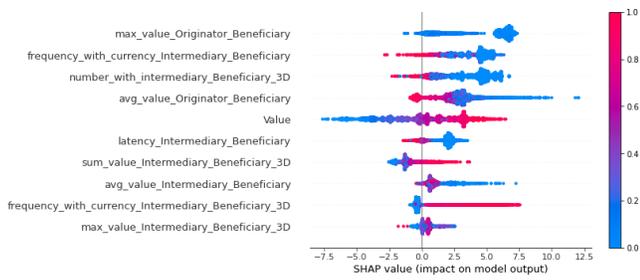


Fig. 4. Beeswarm cluster 0

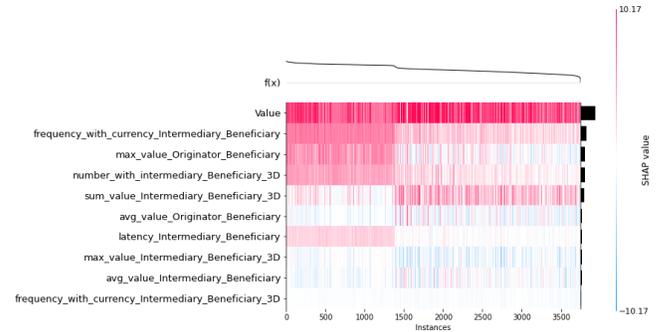


Fig. 7. Heatmap cluster 2

has a high impact with high value, as shown in the *beeswarm* (Fig. 6). (ii) the *heatmap*'s left side has a high impact on the second and third features with high values according to the *beeswarm*. The amount has an average impact corresponding to a medium value for these impacts.

Experts associated this fraud type with customers realizing many transactions (high frequency) in a short time (three days) and with an average amount. This cluster is assigned to the “smurf” fraud type (SF).

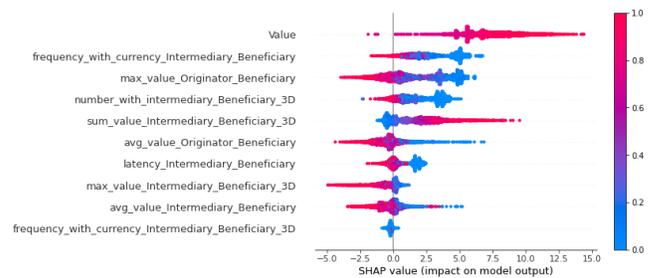


Fig. 8. Beeswarm cluster 2

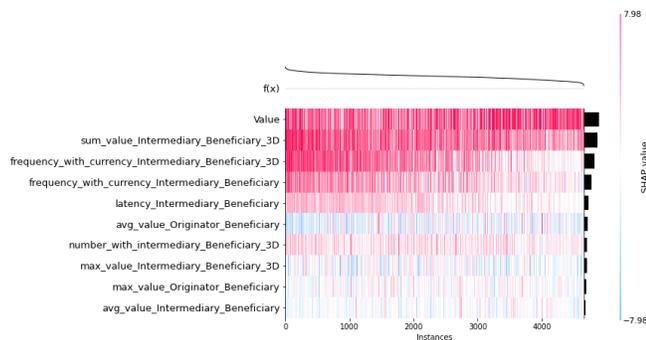


Fig. 5. Heatmap cluster 1

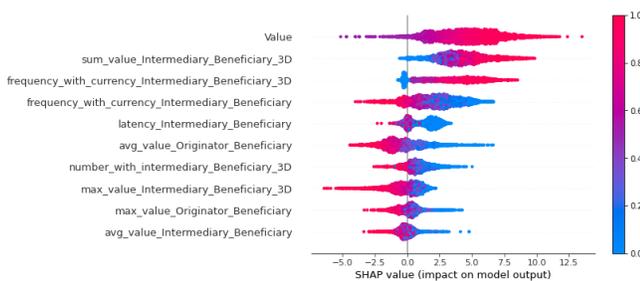


Fig. 6. Beeswarm cluster 1

3) *Cluster 2*: According to the *heatmap* (Fig. 7), the feature *value* (Amount field) greatly impacts the whole cluster. Furthermore, the *beeswarm* (Fig. 8) indicates that its value is high, and if we revert to Table X, the average amount of this cluster is very high. Experts associate this fraud type with customers realizing a high amount of transactions. This cluster is assigned to the “large amount” fraud type (LAF).

We identified three fraud types by leveraging the binary classification model in combination with the SHAP framework and *k-means* algorithm. The next step is training a multi-fraud

classification model to classify frauds in their fraud types. we update the fraudulent transaction label with their fraud type.

### E. Multi-Fraud Classification and Mounting Cost Minimization

We train a multi-Fraud classification model with the same transactions in the training and testing set. Once our model is trained, we choose a threshold for the *MFD* algorithm presented in Section III-E.

we used our cost minimization algorithm to select the optimal threshold. The administrative cost  $C_a$  is set to 100, and the dissatisfaction cost  $C_d$  set to 50.

Fig. 9 shows the model mounting cost, and its f1-score is represented by two curves for each threshold. The mounting cost is low when the threshold is low because transactions probability to belong to fraudulent class are above 0.1. There is no cost impact on the model with fraudulent transaction amounts. However, the model generates many false alerts, for this reason, it's important to maintain a good f1-score. When the threshold is above 0.1, the model has the highest f1-score (0.85). Using our cost minimization algorithm, we retrieve 0.19 as the optimal threshold that minimizes the model mounting cost while maintaining a good f1-score (0.85). We reduce the f1-score of the hundredth order, which is insignificant.

We detail the model's evaluation results with the threshold in Table XII and in the confusion matrix in Table XIII. Our model has a good precision and a weaker recall. It results in a f1-score of 0.74, which is lower than 0.85 with the binary classification. *MFD* algorithm sums the probability of 3 fraud types. The transaction is fraudulent if the sum is above a threshold and its class is the fraud type with

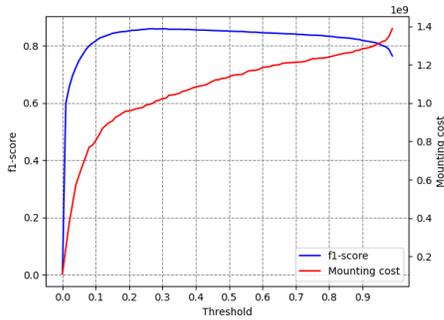


Fig. 9. Mounting costs and f1-score curves

TABLE XII. MULTI-CLASS MODEL

	Fraud Type	Precision	Recall	f1-score
Multi-Class	Legit	0.99	0.99	0.99
	SF	0.91	0.43	0.58
	DAF	0.93	0.50	0.65
	LAF	0.92	0.60	0.73
	Macro avg	0.94	0.63	<b>0.74</b>
Binary (DAF+SF+LAF)	Legit	0.99	0.99	0.99
	Fraud	0.92	0.57	0.70
	Macro avg	0.96	0.78	<b>0.85</b>

the highest probability. In conclusion, our multi-class model detects fraudulent transactions as a binary model, and the fraud type assignment reduces the f1-score of just 0.09 but give additional information to experts about frauds.

V. LIMITS AND DISCUSSIONS

In this section, we completed the 3 goals outlined in section III through experimentation. First, we detect fraudulent transactions by extracting relevant features and using the *CatBoost* algorithm. This resulted in a f1-score of 0.89. However, due to data privacy limitations, we can't compare these results with other datasets. Second, we identify our dataset's fraud types based on the *CatBoost* model, *k-means* algorithm and SHAP framework. A multi-class classification model is trained using the *MFD* algorithm to detect fraudulent transactions and assign them a fraud type. Finally, we select a threshold to detect fraudulent transactions in order to achieve the highest f1-score while minimizing costs. Our multi-class model obtain 0.74 as f1-score (Table XII), which is lower than the 0.89 from the first step. However, if we consider a transaction as fraudulent when it's part on one the 3 fraud types, there is a f1-score of 0.85 (binary row). It means that fraudulent transactions are still detected on this model, however the fraud type assignment is decreasing the f1-score. It can be explained by possible close boundaries between clusters.

TABLE XIII. CONFUSION MATRIX

		Predicted			
		Legitimate	DAF	SF	LAF
Actual	Legitimate	732605	18	2	89
	DAF	317	243	2	2
	SF	55	2	59	0
	LAF	779	2	0	1184

VI. CONCLUSION

We presented in this work the context of financial transactions between banks through the SWIFT network. We studied financial fraud detection literature. Machine learning techniques are valuable for identifying the fraudulent pattern provided during the training phase. We proposed a detection and identification frauds approach based on Desrousseau et al. method. We can summarize our approach as: First, we extracted features from the transaction base fields and on the actors, currencies, countries, and transactions path. Second, we applied a supervised algorithm on our labeled dataset and reduced the features number to retain the relevant one. Third, we used the unsupervised algorithm to cluster fraudulent transactions to identify the fraud types by leveraging the SHAP framework and the supervised model. Then, we trained a multi-fraud classification model to assign a class to a transaction with the three identified fraud types: dormant account fraud, smurf fraud, and large amount fraud. To handle this model prediction, we proposed the *MFD* algorithm to classify a transaction as fraudulent or fraudulent with its type. Finally, we proposed another algorithm to minimize the model mounting cost by choosing a threshold from which a transaction is considered fraudulent.

Our semi-supervised methodology is used by financial institutions to understand their dataset. Cluster interpretation needs experts feedback based on visualization tools (Fig. 3-8).

In conclusion, our contributions with the proposed approach are : detecting fraudulent transactions, identifying fraud types and minimizing the mounting cost.

In future work, we plan to explore additional fraud types on different datasets. We also plan to automatize identifying fraud type with semantic analyses based on financial fraud ontology.

ACKNOWLEDGMENT

The authors would like to thank the University of Burgundy and more particularly the laboratory LIB for the scientific help. We also thank SKAIZen Group company for providing the dataset and experts. Finally, we thank the ANRT for supporting this project.

REFERENCES

- [1] D. Jensen, "Prospective assessment of ai technologies for fraud detection: A case study," in *AAAI Workshop on AI Approaches to Fraud Detection and Risk Management*. Citeseer, 1997, pp. 34–38.
- [2] S. Dasgupta and P. Grover, "Critically evaluating swift's strategy as a monopoly in the fintech business," 2019.
- [3] M. Collin, S. Cook, and K. Soramaki, "The impact of anti-money laundering regulation on payment flows: Evidence from swift data," *Center for Global Development Working Paper*, no. 445, 2016.
- [4] J. L. Perols, "Detecting financial statement fraud: Three essays on fraud predictors, multi-classifier combination and fraud detection using data mining," 2008.
- [5] M. Dixon, I. Halperin, and P. Bilokon, *Machine Learning in Finance: From Theory to Practice*. Springer International Publishing, 2020. [Online]. Available: <https://books.google.fr/books?id=Fuw3zQEACAAJ>
- [6] A. Kumar, S. Das, and V. Tyagi, "Anti money laundering detection using naïve bayes classifier," in *2020 IEEE International Conference on Computing, Power and Communication Technologies (GUCON)*. IEEE, 2020, pp. 568–572.

- [7] U. Porwal and S. Mukund, "Credit card fraud detection in e-commerce: An outlier detection approach," *arXiv preprint arXiv:1811.02196*, 2018.
- [8] F. Carcillo, Y.-A. Le Borgne, O. Caelen, Y. Kessaci, F. Oblé, and G. Bontempi, "Combining unsupervised and supervised learning in credit card fraud detection," *Information sciences*, vol. 557, pp. 317–331, 2021.
- [9] A. Roy, J. Sun, R. Mahoney, L. Alonzi, S. Adams, and P. Beling, "Deep learning detecting fraud in credit card transactions," in *2018 Systems and Information Engineering Design Symposium (SIEDS)*. IEEE, 2018, pp. 129–134.
- [10] L. Keyan and Y. Tingting, "An improved support-vector network model for anti-money laundering," in *2011 Fifth International Conference on Management of e-Commerce and e-Government*. IEEE, 2011, pp. 193–196.
- [11] X. Wang and G. Dong, "Research on money laundering detection based on improved minimum spanning tree clustering and its application," in *2009 Second international symposium on knowledge acquisition and modeling*, vol. 2. IEEE, 2009, pp. 62–64.
- [12] R. A. L. Torres and M. Ladeira, "A proposal for online analysis and identification of fraudulent financial transactions," in *2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE, 2020, pp. 240–245.
- [13] D. Almhaithawi, A. Jafar, and M. Aljnidi, "Example-dependent cost-sensitive credit cards fraud detection using smote and bayes minimum risk," *SN Applied Sciences*, vol. 2, no. 9, pp. 1–12, 2020.
- [14] E. A. Lopez-Rojas and S. Axelsson, "Money laundering detection using synthetic data," in *Annual workshop of the Swedish Artificial Intelligence Society (SAIS)*. Linköping University Electronic Press, Linköpings universitet, 2012.
- [15] J. Tang and J. Yin, "Developing an intelligent data discriminating system of anti-money laundering based on svm," in *2005 International conference on machine learning and cybernetics*, vol. 6. IEEE, 2005, pp. 3453–3457.
- [16] K. Michalak and J. Korczak, "Graph mining approach to suspicious transaction detection," in *2011 Federated conference on computer science and information systems (FedCSIS)*. IEEE, 2011, pp. 69–75.
- [17] L.-T. Lv, N. Ji, and J.-L. Zhang, "A rbf neural network model for anti-money laundering," in *2008 International conference on wavelet analysis and pattern recognition*, vol. 1. IEEE, 2008, pp. 209–215.
- [18] D. Varmedja, M. Karanovic, S. Sladojevic, M. Arsenovic, and A. Anderla, "Credit card fraud detection-machine learning methods," in *2019 18th International Symposium INFOTEH-JAHORINA (INFOTEH)*. IEEE, 2019, pp. 1–5.
- [19] B. Bestami Yuksel, S. Bahtiyar, and A. Yilmazer, "Credit card fraud detection with nca dimensionality reduction," in *13th International Conference on Security of Information and Networks*, 2020, pp. 1–7.
- [20] M. Alkhalili, M. H. Qutqut, and F. Almasalha, "Investigation of applying machine learning for watch-list filtering in anti-money laundering," *IEEE Access*, vol. 9, pp. 18481–18496, 2021.
- [21] S. Bhattacharyya, S. Jha, K. Tharakunnel, and J. C. Westland, "Data mining for credit card fraud: A comparative study," *Decision support systems*, vol. 50, no. 3, pp. 602–613, 2011.
- [22] E. L. Paula, M. Ladeira, R. N. Carvalho, and T. Marzagao, "Deep learning anomaly detection as support fraud investigation in brazilian exports and anti-money laundering," in *2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE, 2016, pp. 954–960.
- [23] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.
- [24] E. Badal-Valero, J. A. Alvarez-Jareño, and J. M. Pavía, "Combining benford's law and machine learning to detect money laundering. an actual spanish court case," *Forensic science international*, vol. 282, pp. 24–34, 2018.
- [25] N. F. Ryman-Tubb, P. Krause, and W. Garn, "How artificial intelligence and machine learning research impacts payment card fraud detection: A survey and industry benchmark," *Engineering Applications of Artificial Intelligence*, vol. 76, pp. 130–157, 2018. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0952197618301520>
- [26] K. G. Al-Hashedi and P. Magalingam, "Financial fraud detection applying data mining techniques: A comprehensive review from 2009 to 2019," *Computer Science Review*, vol. 40, p. 100402, 2021. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1574013721000423>
- [27] M. Albashrawi, "Detecting financial fraud using data mining techniques: A decade review from 2004 to 2015," *Journal of Data Science*, vol. 14, pp. 553–570, 07 2016.
- [28] Y. Zhang and P. Trubey, "Machine learning and sampling scheme: An empirical study of money laundering detection," *Computational Economics*, vol. 54, no. 3, pp. 1043–1063, 2019.
- [29] J. Lorenz, M. I. Silva, D. Aparício, J. T. Ascensão, and P. Bizarro, "Machine learning methods to detect money laundering in the bitcoin blockchain in the presence of label scarcity," in *Proceedings of the First ACM International Conference on AI in Finance*, 2020, pp. 1–8.
- [30] A. Mehbodniya, I. Alam, S. Pande, R. Neware, K. P. Rane, M. Shabaz, and M. V. Madhavan, "Financial fraud detection in healthcare using machine learning and deep learning techniques," *Security and Communication Networks*, vol. 2021, 2021.
- [31] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.
- [32] Y. Freund, R. Schapire, and N. Abe, "A short introduction to boosting," *Journal-Japanese Society For Artificial Intelligence*, vol. 14, no. 771–780, p. 1612, 1999.
- [33] Y. Sun, M. S. Kamel, A. K. Wong, and Y. Wang, "Cost-sensitive boosting for classification of imbalanced data," *Pattern recognition*, vol. 40, no. 12, pp. 3358–3378, 2007.
- [34] A. S. Larik and S. Haider, "Clustering based anomalous transaction reporting," *Procedia Computer Science*, vol. 3, pp. 606–610, 2011.
- [35] J. Guo, G. Liu, Y. Zuo, and J. Wu, "An anomaly detection framework based on autoencoder and nearest neighbor," in *2018 15th International Conference on Service Systems and Service Management (ICSSSM)*. IEEE, 2018, pp. 1–6.
- [36] S. Ramaswamy, R. Rastogi, and K. Shim, "Efficient algorithms for mining outliers from large data sets," in *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, 2000, pp. 427–438.
- [37] N. A. Le Khac and M.-T. Kechadi, "Application of data mining for anti-money laundering detection: A case study," in *2010 IEEE International Conference on Data Mining Workshops*. IEEE, 2010, pp. 577–584.
- [38] S. Raza and S. Haider, "Suspicious activity reporting using dynamic bayesian networks," *Procedia Computer Science*, vol. 3, pp. 987–991, 2011.
- [39] T. Pourhabibi, K.-L. Ong, B. H. Kam, and Y. L. Boo, "Fraud detection: A systematic literature review of graph-based anomaly detection approaches," *Decision Support Systems*, vol. 133, p. 113303, 2020.
- [40] A. C. Bahnsen, A. Stojanovic, D. Aouada, and B. Ottersten, "Cost sensitive credit card fraud detection using bayes minimum risk," in *2013 12th international conference on machine learning and applications*, vol. 1. IEEE, 2013, pp. 333–338.
- [41] R. Desrousseaux, G. Bernard, and J.-J. Mariage, "Profiling money laundering with neural networks: a case study on environmental crime detection," in *2021 IEEE 33rd International Conference on Tools with Artificial Intelligence (ICTAI)*. IEEE, 2021, pp. 364–369.
- [42] T. Kohonen, "The self-organizing map," *Proceedings of the IEEE*, vol. 78, no. 9, pp. 1464–1480, 1990.
- [43] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.
- [44] A. Likas, N. Vlassis, and J. J. Verbeek, "The global k-means clustering algorithm," *Pattern recognition*, vol. 36, no. 2, pp. 451–461, 2003.
- [45] T. M. Kodinariya and P. R. Makwana, "Review on determining number of cluster in k-means clustering," *International Journal*, vol. 1, no. 6, pp. 90–95, 2013.
- [46] A. D. Boyer and S. Light, "Dirty money and bad luck: Money laundering in the brokerage context," *Va. L. & Bus. Rev.*, vol. 3, p. 81, 2008.

# Liver Disease Prediction and Classification using Machine Learning Techniques

Srilatha Tokala, Koduru Hajarathaiyah, Sai Ram Praneeth Gunda, Srinivasrao Botla, Lakshmikanth Nalluri,  
Pathipati Nagamanohar, Satish Anamalamudi, Murali Krishna Enduri  
Department of Computer Science and Engineering, SRM University-AP, Amaravati, India

**Abstract**—Recently liver diseases are becoming most lethal disorder in a number of countries. The count of patients with liver disorder has been going up because of alcohol intake, breathing of harmful gases, and consumption of food which is spoiled and drugs. Liver patient data sets are being studied for the purpose of developing classification models to predict liver disorder. This data set was used to implement prediction and classification algorithms which in turn reduces the workload on doctors. In this work, we proposed apply machine learning algorithms to check the entire patient's liver disorder. Chronic liver disorder is defined as a liver disorder that lasts for at least six months. As a result, we will use the percentage of patients who contract the disease as both positive and negative information. We are processing Liver disease percentages with classifiers, and the results are displayed as a confusion matrix. We proposed several classification schemes that can effectively improve classification performance when a training data set is available. Then, using a machine learning classifier, good and bad values are classified. Thus, the outputs of the proposed classification model show accuracy in predicting the result.

**Keywords**—Machine learning algorithms; classification model; classifier; liver disease

## I. INTRODUCTION

The liver is the most imperative structure in a human build. Insulin is broken down by the liver. The liver breaks bilirubin with glucuronidation, which further helps its defecation into bile [1]. It is also accountable for the breaking down and excretion of many unwanted products. It shows a noteworthy role in altering toxic materials. It shows a noteworthy role in collapsing medicinal products. It's named Drug metabolism. The weight would be 1.3 kg. The liver consists of 2 immense portions namely the privileged portion, and the left estimate. The gallbladder is located below the liver, near the pancreas. The Liver along with these organs helps to consume and give nutrition. Its job is to help the flow of the wounding materials in the stream of blood from the stomach, before passing it to whatsoever is left of the body. Liver sicknesses are triggered when the working of the liver is affected or any injury has happened to it [2].

The development of liver disorders [3] is complicated and varied in character, influenced by a number of variables that determine disease susceptibility. Sex, ethnicity, genetics, environmental exposures (viruses, alcohol, nutrition, and chemicals), body mass index (BMI), and coexisting diseases like diabetes are among them. A high mortality rate is associated with liver problems, which are life-threatening diseases. The usual urine and blood tests are the first step in the prognosis of liver disorders. A LFT (liver functions test) is recommended for the patient based on the symptoms seen [4].

Liver disease is a significant health issue affecting millions of people globally. Early detection and accurate classification of liver diseases can lead to better patient outcomes and reduce the burden on the healthcare system. One-third of adults and an increasing proportion of youngsters in affluent nations suffer from non-alcoholic fatty liver disease (NAFLD) [5], a growing health issue. The abnormal buildup of triglycerides in the liver, which in some people causes an inflammatory reaction that can lead to cirrhosis and liver cancer, is the first sign of the condition. While there is a significant correlation between obesity, insulin resistance, and non-alcoholic fatty liver disease (NAFLD), the pathophysiology of NAFLD remains poorly understood, and treatment options are limited. However, machine learning techniques have demonstrated encouraging results in predicting and categorizing liver diseases based on patient data. By utilizing sophisticated algorithms to analyze and learn from large datasets, these techniques can identify patterns and anticipate outcomes. The employment of machine learning techniques in liver disease prediction and classification is a dynamic area of research, with continual advancements being made to enhance accuracy and decrease healthcare costs.

### A. Overview of Liver Disease

Liver disease refers to an abnormality in the liver's function, resulting in illness [4]. The liver is responsible for many vital functions within the body, and if it becomes damaged or infected, the loss of these functions can have a significant impact on overall health. Hepatic disorder is another term used to describe liver disease [6]. This umbrella term encompasses a range of possible complications that prevent the liver from performing its assigned roles. Even if only a quarter of the liver is still functioning and the rest is damaged, this organ's efficiency will be greatly reduced. The liver is the biggest hard structure in the human build and is well thought-out as a gland because, amid its many roles, it creates and secretes bile. The liver is stood at the upright part of the abdomen and the rib cage shelters it. It has two core lobes that are thru with small lobules. The liver cells have two dissimilar bases of a blood source. The hepatic artery transfers heart-driven blood abundant in oxygen, while the portal vein provides nutrients from the intestines. Generally, the vein's job is to bring the blood from all other organs to the heart, but the portal vein permits nutrients from the digestive region to go into the liver for treating and purifying the former to flow into the general circulation. The portal vein proficiently transports the chemicals that liver cells require to yield the proteins, cholesterol, and glycogen needed for usual body actions.

## B. Causes of Liver Disease

There are numerous activities that prompt liver maladies [7]. The classifications are:

1) *Infection*: The liver can become infected by parasites and viruses, which can lead to inflammation or edoema and compromise liver function. The virus that typically results in liver damage is spread through blood or sperm and is primarily brought on by tainted food, contaminated water, or contact with an infected person. Hepatitis A, C, and B are liver infections that can affect people.

2) *Immune system abnormality*: The body's immune system is administered by certain ailments, to attack other body parts. The liver is also affected. These diseases could be Autoimmune hepatitis. In addition, it could be Primary biliary cholangitis, and Primary sclerosing cholangitis.

3) *Inheritance*: A rare gene genetically inherited from either of your parents can cause a buildup of various compounds in the liver, which can cause liver damage. Wilson's disease, Hemochromatosis, and alpha-1 antitrypsin deficiency are three examples of genetic liver illnesses.

4) *Cancer and other progressions*: Cancers that have may reason liver diseases are Liver adenoma, Bile duct cancer, and Liver cancer.

5) *Others*: The general reasons are prolonged alcohol abuse, fat buildup in the liver (NAFLD), certain drugs or over-the-counter treatments, and certain herbal mixes.

6) *Risk aspects*: Factors that might raise the risk of liver diseases are excessive usage of liquor, being overweight, diabetes of type, tattoos, piercings of the body, drug injection with used needles, transfusion of blood, exposure to foreign blood, unprotected intercourse, exposure to chemicals, and inheritance.

## C. Chemicals Compounds in Liver

Chemicals such as Bilirubin, Albumin, Alkaline phosphatase, Aspartate aminotransferase, and globulin are existent in the liver and perform a vital role in the daily operations of the healthy liver.

1) *Bilirubin*: Bilirubin is a yellowish complex that arises in the usual catabolic trail that breaks down heme in vertebrates. Bile and urine emit it. Raised volumes of bilirubin in the body cause diseases. The bilirubin is accountable for the yellow shade of cuts and the yellow staining in jaundice disease. Its following breakdown products, like stercobilin, are accountable for the brown color of feces. Another breakdown product, urobilin, is the key constituent of the straw-yellow color of urine.

2) *Alkaline phosphatase*: In beings, alkaline phosphatase is existent in all tissues all over the body but is mainly focused in the liver, intestinal mucosa, bile duct, bone, kidney, and placenta. In the serum, two kinds of alkaline phosphatase isozymes prevail skeletal and liver. In childhood, most of the alkaline phosphatase is of the skeletal source. Most of the mammals including humans have these types of alkaline phosphatases:

- ALPI: It is intestinal having a molecular mass of 150 kDa.
- ALPL: It is tissue-nonspecific mainly present in the liver, kidney, and bone.
- ALPP: It is placental and is also known as Regan isozyme.
- GCAP: It is a germ cell.

3) *Aspartate aminotransferase*: AST is a kind of enzyme. AST levels are higher in the heart and liver. AST is found in the kidneys and muscles, although in less amounts. It is very low in human blood. When muscle or liver cells are injured, the AST is released into the bloodstream. The AST test will therefore be useful for tracking or identifying liver damage or dysfunction.

4) *Albumin*: They're globular proteins. Serum albumins are common and are the most imperative protein of blood. It binds thyroxine (T4), water, cations like Ca<sup>2+</sup> and Na<sup>+</sup>, hormones, fatty acids, bilirubin, and pharmaceuticals. Its core part is to govern and normalize the oncotic pressure of the blood. It binds several fatty acids, cations, and bilirubin.

5) *Globulin*: They are protein globules. They are heavier than albumin at the molecular level. It will not dissolve in pure water but will solvate in dilute salt solutions. The liver produces some globulins. Globulin absorption in fit human blood is around 2.6-3.5 g/dL. There are several different types of globulins, including beta, alpha 1, alpha 2, and gamma globulins. Any unfitting amounts of these chemicals produced in the kidney can reason an imbalance and cause liver diseases. These are considered features. There are  $n$  number of kinds of liver illnesses and these are grounded based on the proportion of these chemicals stashed.

## II. MACHINE LEARNING AND LIVER DISEASE

Classification algorithms are often defined to be used for forecasting the liver disease as they can help predict whether a patient has the disease or not based on certain features or characteristics. Based on the existing solutions, it was found that the F-Tree algorithm shows the highest accuracy among the algorithms tested, making it a suitable choice for forecasting liver disease. Feature selection along with the fuzzy K-means classification methods are commonly used in the classification of liver diseases. These methods can help identify important features that can be used to distinguish between different types of liver disorders. As the same attribute values may be present in different liver disorders, using fuzzy-based classification can help improve the performance of the classification process by taking into account the degree of similarity between instances [8], several classification algorithms, including J48, SVM, RF, and MLP, were used to classify liver diseases. The study evaluated the performance of these state-of-the-art algorithms using metrics such as data accuracy, data effectiveness, and correction rate, and compared the results. The findings indicated that the multilayer perceptron algorithm achieved the highest accuracy compared to the other algorithms examined in the research. In this work [9], researchers employed Bayesian classification to distinguish between various liver diseases, including cirrhosis, hepatitis, and non-liver diseases. They used both Naive Bayes and FT

tree techniques to categorize the liver patient dataset into different illness subtypes, and evaluated the performance of these methods in terms of accuracy and execution time. Their analysis revealed that Naive Bayes algorithm outperformed the FT tree algorithm in terms of execution time.

In this work [10], authors explored the effectiveness of several classification methods in diagnosing liver diseases. These methods included Naive Bayes, Decision Tree, Multilayer Perceptron, K-Nearest Neighbors, Random Forest, and Logistic Regression. To evaluate their performance, the authors used metrics such as precision, recall, sensitivity, and specificity. The results showed that Naive Bayes achieved the highest precision compared to the other algorithms examined. Additionally, the Logistic Regression and Random Forest algorithms were found to have good results when recall was considered.

In [11], to create a model for analysing liver illness, the WEKA tool was employed. To create their suggested model, the Naive Bayes, Decision Tree (DT), and *J48* algorithms were employed. The algorithms' accuracy and execution time were measured, and the results were compared to those of the available options. The findings demonstrated that the *J48* and DT algorithms outperformed the Naive Bayes algorithm in terms of accuracy.

In [12], uses a dataset of Indian patients with liver disease and implements it using various classification algorithms, including LR, K-NN, and SVM. Additionally, confusion matrix was used to evaluate the algorithms against one another. Based on their correctness, the experimental analysis of these algorithms was examined and evaluated. According to the findings, LR and KNN techniques can both achieve a sizable degree of accuracy, although LR has a high sensitivity. With this, it can be inferred that LR is an appropriate strategy for disease prediction. A novel feature model with the classification methods Random Forest, SVM, *J48*, Bayesian Net, and MLP was proposed in [13], [14], [15].

Three distinct procedures are used to examine and implement a comparative study for the prediction of liver disease. On the dataset containing records of liver diseases, the normalisation method min-max was examined and tested in the initial phase. The second step uses the PSO feature selection approach to choose the dataset's necessary components for predicting liver illness. The implementation of the classification algorithms and an accuracy-based performance evaluation of the methods were done in the third phase. The experimental study leads to the conclusion that the *J48* method outperformed when PSO feature selection was used. In this work [16], aimed to balance the dataset for accurate prediction of liver diseases by using a combination of sampling and oversampling techniques. Several classification algorithms were used, including *J48*, Multilayer perceptron, Random Forest, Multiple linear regression, Support Vector Machines, and Genetic programming. The Random Forest algorithm with oversampling at higher rates demonstrated the best performance in predicting liver diseases. Similarly, In this work [17] utilized various classification algorithms such as Naive Bayes, FT tree, *J48*, SVM, RF, MLP, K-NN, LR, and *C4.5* to classify liver diseases. Different techniques such as feature selection, fuzzy K-means classification, normalization, PSO feature selection, oversampling, and undersampling were

also employed to enhance the performance of these algorithms. The performance of these algorithms was evaluated based on different factors such as accuracy, execution time, precision, recall, sensitivity, specificity, and F-measures. The results of these studies indicate that different algorithms perform differently depending on the dataset and application [18]. Therefore, it is crucial to evaluate the performance of multiple algorithms and techniques to identify the best approach for a specific dataset and application. Although the outcomes have been found to differ depending on the dataset employed and the specific algorithm used, some algorithms such as Multilayer perceptron, Random Forest, Support Vector Machines, and Artificial Neural Network generally perform better than others in predicting liver diseases.

ML has a multiplied widespread routine in the modern ages. Using ML as a means of aid in therapeutic and pharmacological diagnostics is increasing. However, the common thing is rapidly increasing access to huge amounts of data. The following content gives a thorough summary of the major defies and prevalent resolutions to ML challenges [19] in medicine. Feature selection is to decrease the struggle by selecting a subsection of the useful features in the input and disposing of the residual features.

It is clear from the above statement that various studies have been conducted using different machine learning algorithms to predict liver diseases and evaluate their performance. Different algorithms such as Naive Bayes, Decision Trees, Random Forest, Logistic Regression, K-NN, SVM, Artificial Neural Networks, and *C4.5* have been used in these studies. The performance of these algorithms has been evaluated based on factors such as accuracy, precision, recall, F1-score, specificity, and execution time. It is also noted that different studies have different conclusion on which algorithm performs best, it depends on the dataset and evaluation metrics used.

#### A. Logistic Regression

Its new name is the Logit model [20]. It is used to simulate the likelihood of a specific class or event prevailing, such as pass/fail, healthy/sick, alive/dead, or win/lose. It's a mathematical model. In its most basic form, it employs a logistic strategy to advance a binary dependent variable, while several complicated expansions exist. A binary logistic model [21] will include a variable that is dependent and has two possible values, such as pass/fail, which represents an indicator variable, and the two values are regarded as 0 and 1.

#### B. Support Vector Machine

It targets to find an ideal hyperplane that splits the data into diverse classes. A way of implementing SVM [22] in python is by making use of the scikit-learn package. The data is sorted out and is disjointed into test data and training sets. The testing data is set as twenty out of every hundred. The set for algorithm preparation is fixed as one hundred and sixty out of every two hundred. An SVM forms either a set of hyperplanes. It is built in an immeasurable-extension space. A hyperplane that has the farthest distance to the nearest data point of any class achieves a good separation, called a functional margin. In essence, a wider margin implies a smaller simplification defect for the classifier.

### C. Convolutional Neural Networks

CNN is a portion of DNN [23]. It is broadly purposed for the exploration of imageries. They are also known as shift invariant ANN or space invariant ANN. They have a mutual-bulks architecture. Here, the network develops a mathematical process termed convolution. They are solely, neural networks that practice convolution as a replacement for general matrix multiplication in a minimum of one of their layers.

### D. MLP Classifier

MLP [24] is a portion of (ANN). Also called, feed-forward ANN. An MLP encompasses the bottom of three sheets of nodes. The three sheets are the first input, the mid-hidden, and the final output layers. Further, each other node is doubled as a neuron that makes use of a nonlinear activation, with an exception of the input node. MLP practices an administered learning system that is known as backpropagation, for training. MLP [25] can be distinguished from linear perceptron because of its manifold layers and non-linear instigation.

### E. Random Forest

The random forest [26] creates decision trees on data mockups and gains the prediction from all the formed decision trees. Finally, it elects the best elucidation by voting means. The data is sorted out and is disjointed into test data and training sets. The testing data is set as twenty out of every hundred. The set for algorithm preparation is fixed as one hundred and sixty out of every two hundred. The program fragments the data into many groups and subgroups. If an individual draws lines between data points in a subgroup, lines connecting subgroups into a group, and so on. The erection appears to be tree-like. The hyperplane that maximizes the distance to the nearest data point in the training set provides a reliable separation between classes.

## III. LITERATURE REVIEW

### A. Liver Disease Biopsies using Deep Learning and CNN

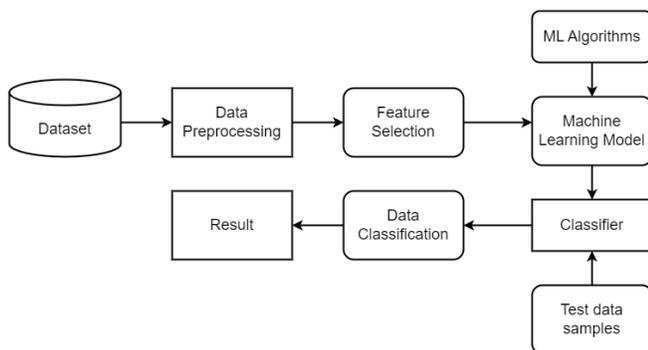


Fig. 1. Processing of ML algorithms

The author sought to implement a completely involuntary tool for diagnosing liver disease by using liver biopsy images. The author considered using biopsy images because there is a respectable chance of differentiating an unhealthy and healthy liver using these images. The projected tactic is to

use image study and deep learning and further determine an efficient CNN [27] architecture and further execution. NAFLD is common. An investigation stated that almost forty percent of all liver illnesses around the globe are caused by NAFLD. The existence of NAFLD in the liver can be found by the sign of hepatic steatosis, and also other reasons for fat build-up, such as major consumption of alcohol, lasting use of steatogenic medicines, and genetic problems. In this proposed method, the biopsy images of the liver are taken and the hepatic structures existent in these are analyzed with the aid of two CNN [28] which have the same architecture. They want to develop a 4-class detection system, which detects sinusoids, ballooned hepatocytes, veins, and fat droplets initially. In the concluding phase, this detection system is united with each other to complete the methodology. It then calculates the fat and ballooning ratio. The found ratio will aid in concluding the patient's condition. They made use of seven hundred and twenty liver biopsy images. Six hundred and twenty of these images are used for testing the algorithm, sixty are used for validation and the rest of the forty images are used for testing. These images originally are of 10,000 \* 10,000 pixels or above. The area in which the tissue is extant is hauled out from each image and the resultant images are 64 \* 64 pixels.

### B. Predicting the Accuracy of Liver Disease using Machine Learning

Utilizing and modeling medicinal data sets are now considered by specialists across the globe. The main plan here is to shorten the time gap in the middle of testing the liver, and generating the report and final result. They used some Machine Learning algorithms [29] like decision tree, naive Bayes, ANN, and random forest. Then, the Pearson correlation [30] to find the anomalies such as TP (true positives), FN (false negatives), FP (false positives), and TN (true negatives) is applied. This is done to find the precision, specificity, and affectability of the algorithms used. The produced words are used to compute the sensitivity, affectability, specificity, and accuracy using pre-defined equations. The author intended to create an interface in which the user could enter the patient's report as input. The algorithms are then skill-trained using the allocated data set, and the output of the user input information is determined. The output will be a single number that is either a zero or a one, with the binary one indicating that the patient's liver is sickly and the binary zero indicating that the patient's liver is healthy. The user-given input data will be logged and this data will further train the algorithm again. These additional tabulated values will train and skill the algorithm to progress with precision. UCI machine is the site from where the authors obtained the data set. The outcoming results of these respective algorithms are charted. These grids show outcomes, which are encouraging Accuracy not cited. They made use of ML algorithms- ANN, Naïve Bayes, and Decision tree to make the model. There are numerous types of liver ailments, the authors considered the general liver diseases to ease the process and get a precise result.

### C. Segmentation of Liver using CT Scan and Finding Disease

The authors lit up an idea by making the use of Abdominal CT, liver disease can be perceived. Some organs cannot be perceived through standard X-Ray equipment. These are the

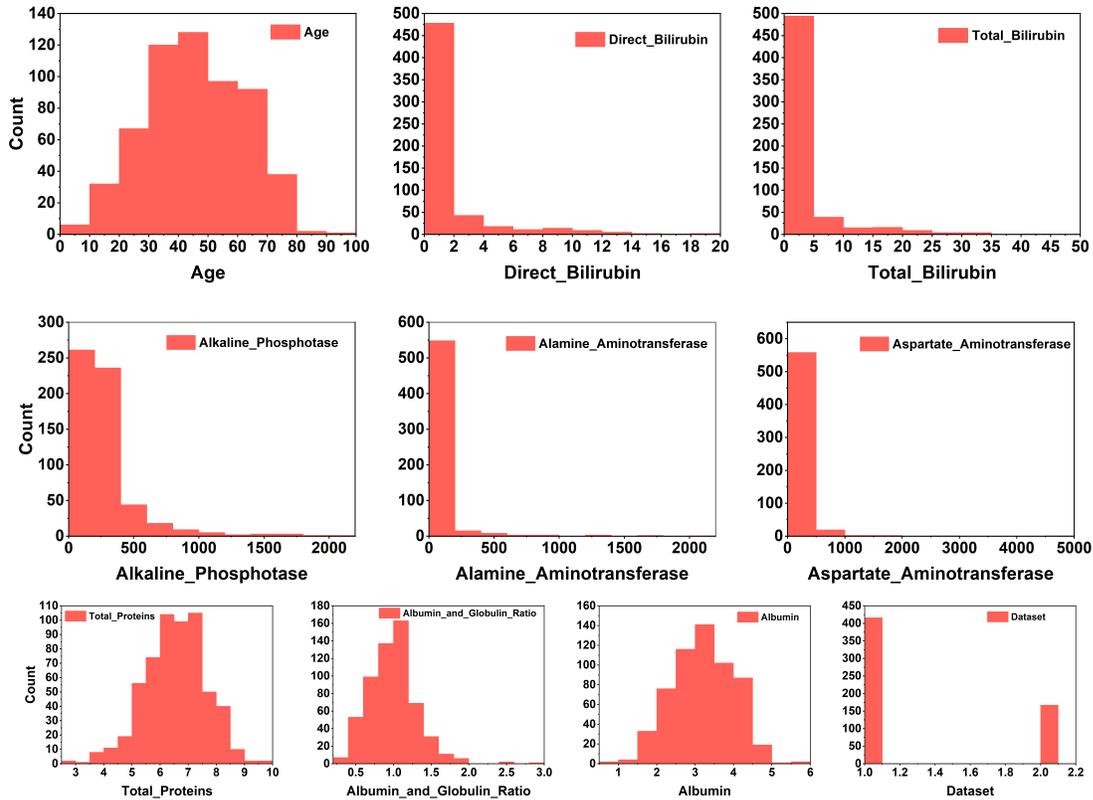


Fig. 2. Histogram for frequency distributions of various patients based on attribute

TABLE I. ACCURACY FOR ML METHODS

ML Models	Accuracy
Logistic Regression	0.76
Support Vector Machine	0.72
Nearest Neighborhood	0.80
Random Forest	0.88

conditions that strengthen the motives to use CT Scans as they can show the structures better than an X-Ray. These CT scan-produced images will have an accurate resolution. They proposed to use WTA to segment the Scan image, identify the liver placement, and differentiate it from the background. In the ending step, the percentage of the area affected is calculated. With the intention of perceiving the progress or sternness of the liver disease, one should use a highly precise technique, that is CT Scan. This CT Scan is extensively used in this medical field to attain info about the humanoid build. In the initial step, the imageries are scrutinized to find different parts. In the following step, the images are handled using the Erode and Dilate algorithm. Viewing point values are adjusted here. Further, they are processed with WTA to segment the liver area. The WTA gives two outputs namely the liver area and the non-liver area. The output yielded is referred to as cropping. Then the gathered copy is adapted into a binary where this organ is white and the rest is black. Then median filtering is done to reduce the noise and smoothens the texture. Then the damaged region of the liver is tallied using a formula. The authors affirm that the typical precision of image breakdown

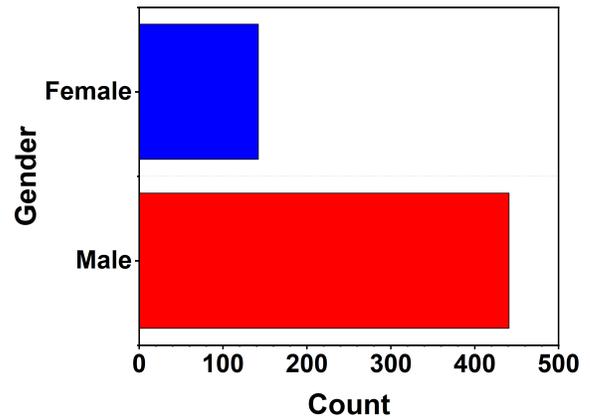


Fig. 3. Comparison with the frequency of males and females

is around eighty-one percent and the typical precision of liver breakdown is around 92 percent. The author used the WTA to distinguish the liver. Another method suggested is to use the binary threshold to isolate the liver area and the diseased area. The closing process in this paper is to measure the fraction of diseased spaces in the liver.

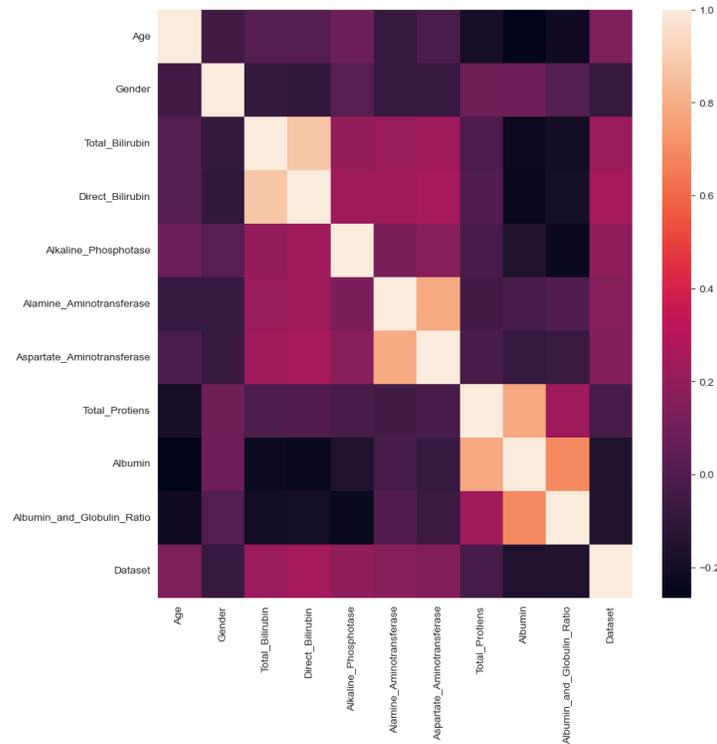


Fig. 4. The correlation graph between each attribute

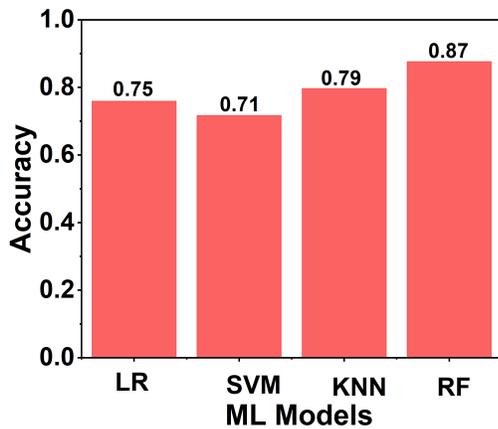


Fig. 5. The accuracy of each machine learning technique used for classification is compared with others

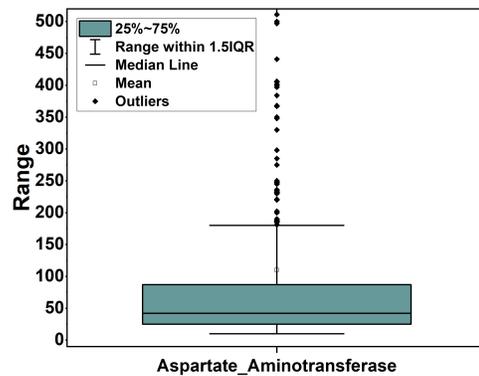


Fig. 6. The box plot represents about the outliers present in that attribute

#### D. Liver Disease Prediction using Classification Algorithms

The proposed process by the authors shows that machine learning can not only predict the disease but also can recognize hidden patterns for diagnosis and decision-making. The everyday growing cases of liver disorders are considered to be a common problem around the world. The goal of this thesis is to provide competent findings in identifying liver disease using classification algorithms [31]. Logistic Regression, K-Nearest Neighbor, and Support Vector Machines are the algorithms utilized for this type of job. The algorithms called Classification

algorithms [32] are predominantly being used in predicting diseases among machine learning algorithms. ML techniques are now very helpful in the healthcare sector for the prediction of diseases from medical databases. In almost every continent there are researchers and scientists who are rigorously using machine learning models [33] with classification algorithms to strategically enhance medical diagnostics and are showing better results. Logistic Regression, K-Nearest Neighbor, and Support Vector Machines are used in this paper to predict liver disease. We all know that the liver is the body’s largest internal organ, performing vital functions such as producing blood clotting factors and proteins, producing triglycerides and

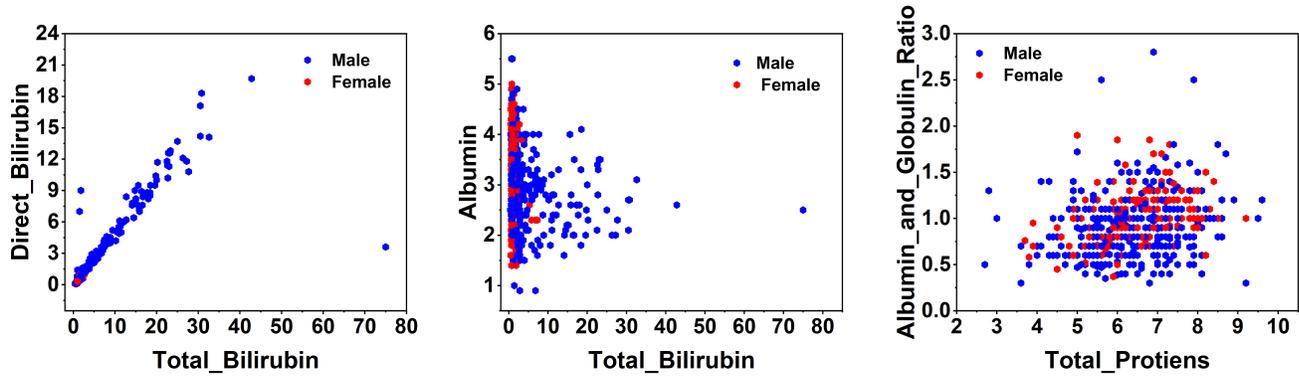


Fig. 7. Scatter plots for Direct\_Bilirubin vs Total\_Bilirubin, Albumin vs Total\_Bilirubin, Albumin\_and\_Globulin\_Ratio vs Total\_Proteins

cholesterol, glycogen synthesis, and bile production. In most cases, a decrease in function must affect much more than 75% of the liver tissue. So it's crucial to notice at an untimely stage at which the disease can be treated before it progresses to the severe stage

#### IV. PROPOSED SYSTEM

In the proposed system, we have to import the liver patient dataset (.csv). Then the dataset is pre-processed and the anomalies and full-up empty cells in the dataset are removed, so that we can further improve the effective liver disease prediction. Then we construct a Confusion matrix for accomplishing an enhanced lucidity of the no of correct/incorrect predictions. Formerly, several classification and prediction procedures and if possible, combinations of different algorithms are implemented and check the accuracy. Our objective is to develop a code that delivers an exactitude of 90%. The advantages are improved classification, early prediction of risks, and improved accuracy. The block diagram of the overall system is shown in Fig. 1.

#### V. RESULTS

Using various methods, we begin our study in this part with the data-processing stage and go on to feature extraction, classification, and prediction analysis. The attributes used in the dataset are Age, Direct Bilirubin, Total Bilirubin, Alkaline Phosphate, Alamine Aminotransferase, Aspartate Aminotransferase, Total Proteins, Albumin, and Globulin Ratio, Albumin, Dataset (where data set is the class label). Each histogram tells us about the frequency distributions for various patients in that particular attribute, shown in Fig. 2.

We represented attributes count or frequency of the patients based on the age, direct\_bilirubin, total\_bilirubin, alkaline\_phosphotase, alamine\_aminotransferase, aspartate\_aminotransferase, total\_proteins, albumin\_and\_globulin\_ratio, albumin, and data set. The Fig 4 shows the correlation between each attribute used in data set are plotted. The lighter the color between two attributes in the graph the higher the values of one attribute are dependent or correlated on the second attribute. From the Fig. 4 we can say that direct\_bilirubin and total\_bilirubin are highly correlated.

The Fig. 3 the frequency of males used in the dataset is compared with frequency of females. The Data set contains a total of 441 males and 142 females. The accuracy of each model is obtained by training the model with the dataset values and testing it by predicting the dataset value. The number of correct predictions done by the model gives us accuracy. From Fig. 5, we can say that Random Forest has the highest accuracy compared to other models. The box plot is plotted which tells us about the outliers present in that attribute. The box plot identifies the outliers using IQR (Inter Quartile Range) method. From Fig. 6, we can see only two values are greater than 450 and far from other values. Hence they are outliers. From Fig. 7 scatter plots for Direct\_Bilirubin vs Total Bilirubin, Albumin vs Total\_Bilirubin, Albumin and Globulin Ratio vs Total Proteins are plotted. Scatter plots are a valuable tool for visualizing the relationship between variables, with dots representing the data points. They are commonly employed to illustrate the associations between variables and how changes in one variable impact another. From the Fig. 7, we can see for Direct\_Bilirubin vs Total Bilirubin the scatter plot is like a straight line which indicates they are highly related. Table I shows the accuracy of the ML method results. The random forest method gives good accuracy than LR, SVM, and nearest neighborhood.

#### VI. CONCLUSION

The liver patient data set was used to implement prediction and classification algorithms, which in turn reduces the workload on doctors. We suggested employing machine learning techniques to examine the patient's total liver condition. A liver condition that has persisted for at least six months is considered chronic. We will thus utilise the proportion of people who get the condition as both positive and negative data. A confusion matrix is used to represent the outcomes of classifier processing of percentages of liver disease. When a training data set is available, our proposed classification schemes can significantly enhance classification performance. Then, using a machine learning classifier, good and bad values are classified. Thus, the outputs of the proposed classification model show accuracy in predicting the result.

The extent of our work is that we will apply deep learning techniques to predict liver disease. Some of the future directions are improve the accuracy of liver disease prediction

and classification models is to include more diverse data sources, improving liver disease prediction and classification is to combine multiple machine learning techniques, machine learning models can be trained to predict the likelihood of liver disease in individuals based on their unique characteristics. Another important direction in liver disease prediction and classification using machine learning is to develop models that are explainable. This means that the models should provide clear and interpretable insights into the factors that contribute to liver disease. Explainable models can help healthcare professionals to make better decisions and provide better care for patients.

## REFERENCES

- [1] A. Arjmand, C. T. Angelis, A. T. Tzallas, M. G. Tsiouras, E. Glavas, R. Forlano, P. Manousou, and N. Giannakeas, "Deep learning in liver biopsies using convolutional neural networks," in *2019 42nd International Conference on Telecommunications and Signal Processing (TSP)*. IEEE, 2019, pp. 496–499.
- [2] L. A. Auxilia, "Accuracy prediction using machine learning techniques for indian patient liver disease," in *2018 2nd International Conference on Trends in Electronics and Informatics (ICOEI)*. IEEE, 2018, pp. 45–50.
- [3] A. Spann, A. Yasodhara, J. Kang, K. Watt, B. Wang, A. Goldenberg, and M. Bhat, "Applying machine learning in liver disease and transplantation: a comprehensive review," *Hepatology*, vol. 71, no. 3, pp. 1093–1105, 2020.
- [4] S. Sontakke, J. Lohokare, and R. Dani, "Diagnosis of liver diseases using machine learning," in *2017 International Conference on Emerging Trends & Innovation in ICT (ICEI)*. IEEE, 2017, pp. 129–133.
- [5] J. C. Cohen, J. D. Horton, and H. H. Hobbs, "Human fatty liver disease: old questions and new insights," *Science*, vol. 332, no. 6037, pp. 1519–1523, 2011.
- [6] F. Himmah, R. Sigit, and T. Harsono, "Segmentation of liver using abdominal ct scan to detection liver disease area," in *2018 International Electronics Symposium on Knowledge Creation and Intelligent Computing (IES-KCIC)*. IEEE, 2018, pp. 225–228.
- [7] M. B. Priya, P. L. Juliet, and P. Tamilselvi, "Performance analysis of liver disease prediction using machine learning algorithms," *International Research Journal of Engineering and Technology (IRJET)*, vol. 5, no. 1, pp. 206–211, 2018.
- [8] T. R. Baitharu and S. K. Pani, "Analysis of data mining techniques for healthcare decision support system using liver disorder dataset," *Procedia Computer Science*, vol. 85, pp. 862–870, 2016.
- [9] U. R. Acharya, S. V. Sree, R. Ribeiro, G. Krishnamurthi, R. T. Marinho, J. Sanches, and J. S. Suri, "Data mining framework for fatty liver disease classification in ultrasound: a hybrid feature extraction paradigm," *Medical physics*, vol. 39, no. 7Part1, pp. 4255–4264, 2012.
- [10] N. Nahar and F. Ara, "Liver disease prediction by using different decision tree techniques," *International Journal of Data Mining & Knowledge Management Process*, vol. 8, no. 2, pp. 01–09, 2018.
- [11] A. Naik and L. Samant, "Correlation review of classification algorithm using data mining tool: Weka, rapidminer, tanagra, orange and knime," *Procedia Computer Science*, vol. 85, pp. 662–668, 2016.
- [12] A. N. Arbain and B. Y. P. Balakrishnan, "A comparison of data mining algorithms for liver disease prediction on imbalanced data," *International Journal of Data Science and Advanced Analytics (ISSN 2563-4429)*, vol. 1, no. 1, pp. 1–11, 2019.
- [13] M. A. Kuzhippallil, C. Joseph, and A. Kannan, "Comparative analysis of machine learning techniques for indian liver disease patients," in *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)*. IEEE, 2020, pp. 778–782.
- [14] K. R. Asish, A. Gupta, A. Kumar, A. Mason, M. K. Enduri, and S. Anamalamudi, "A tool for fake news detection using machine learning techniques," in *2022 2nd International Conference on Intelligent Technologies (CONIT)*. IEEE, 2022, pp. 1–6.
- [15] M. K. Enduri, A. R. Sangi, S. Anamalamudi, R. C. B. Manikanta, K. Y. Reddy, P. L. Yeswanth, S. K. S. Reddy, and G. A. Karthikeya, "Comparative study on sentimental analysis using machine learning techniques," *Mehran University Research Journal of Engineering and Technology*, vol. 42, no. 1, pp. 207–215, 2023.
- [16] M. Islam, C.-C. Wu, T. N. Poly, H.-C. Yang, Y.-C. J. Li *et al.*, "Applications of machine learning in fatty liver disease prediction," in *Building Continents of Knowledge in Oceans of Data: The Future of Co-Created eHealth*. IOS Press, 2018, pp. 166–170.
- [17] S. Mohanty, P. K. Gantayat, S. Dash, B. P. Mishra, and S. C. Barik, "Liver disease prediction using machine learning algorithm," in *Data Engineering and Intelligent Computing: Proceedings of ICICC 2020*. Springer, 2021, pp. 589–596.
- [18] C. Liang and L. Peng, "An automated diagnosis system of liver disease using artificial immune and genetic algorithms," *JOURNAL OF MEDICAL SYSTEMS*, vol. 37, no. 2, 2013.
- [19] R. A. Khan, Y. Luo, and F.-X. Wu, "Machine learning based liver disease diagnosis: A systematic review," *Neurocomputing*, vol. 468, pp. 492–509, 2022.
- [20] A. S. Abdalrada, O. H. Yahya, A. H. M. Alaidi, N. A. Hussein, H. T. Alrikabi, and T. A.-Q. Al-Quraishi, "A predictive model for liver disease progression based on logistic regression algorithm," *Periodicals of Engineering and Natural Sciences (PEN)*, vol. 7, no. 3, pp. 1255–1264, 2019.
- [21] F. E. Harrell, Jr and F. E. Harrell, "Binary logistic regression," *Regression modeling strategies: With applications to linear models, logistic and ordinal regression, and survival analysis*, pp. 219–274, 2015.
- [22] E. M. Hashem and M. S. Mabrouk, "A study of support vector machine algorithm for liver disease diagnosis," *American Journal of Intelligent Systems*, vol. 4, no. 1, pp. 9–14, 2014.
- [23] Z. Yao, J. Li, Z. Guan, Y. Ye, and Y. Chen, "Liver disease screening based on densely connected deep neural networks," *Neural Networks*, vol. 123, pp. 299–304, 2020.
- [24] M. Abdar, N. Y. Yen, and J. C.-S. Hung, "Improving the diagnosis of liver disease using multilayer perceptron neural network and boosted decision trees," *Journal of Medical and Biological Engineering*, vol. 38, no. 6, pp. 953–965, 2018.
- [25] T. Bikku, "Multi-layered deep learning perceptron approach for health risk prediction," *Journal of Big Data*, vol. 7, no. 1, pp. 1–14, 2020.
- [26] T. A. Assegie, R. Subhashni, N. K. Kumar, J. P. Manivannan, P. Duraisamy, and M. F. Engidaye, "Random forest and support vector machine based hybrid liver disease detection," *Bulletin of Electrical Engineering and Informatics*, vol. 11, no. 3, pp. 1650–1656, 2022.
- [27] J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, T. Liu, X. Wang, G. Wang, J. Cai *et al.*, "Recent advances in convolutional neural networks," *Pattern recognition*, vol. 77, pp. 354–377, 2018.
- [28] J. Murphy, "An overview of convolutional neural network architectures for deep learning," *Microway Inc*, pp. 1–22, 2016.
- [29] G. Bonaccorso, *Machine learning algorithms*. Packt Publishing Ltd, 2017.
- [30] I. Cohen, Y. Huang, J. Chen, J. Benesty, J. Benesty, J. Chen, Y. Huang, and I. Cohen, "Pearson correlation coefficient," *Noise reduction in speech processing*, pp. 1–4, 2009.
- [31] M. Ghosh, M. Raihan, M. Sarker, M. Raihan, L. Akter, A. K. Bairagi, S. S. Alshamrani, and M. Masud, "A comparative analysis of machine learning algorithms to predict liver disease," *Intelligent Automation & Soft Computing*, vol. 30, no. 3, 2021.
- [32] F. Osisanwo, J. Akinsola, O. Awodele, J. Hinmikaiye, O. Olakanmi, and J. Akinjobi, "Supervised machine learning algorithms: classification and comparison," *International Journal of Computer Trends and Technology (IJCTT)*, vol. 48, no. 3, pp. 128–138, 2017.
- [33] P. C. Sen, M. Hajra, and M. Ghosh, "Supervised classification algorithms in machine learning: A survey and review," in *Emerging technology in modelling and graphics*. Springer, 2020, pp. 99–111.

# Image Super-Resolution using Generative Adversarial Networks with EfficientNetV2

Saleh AlTakroui<sup>1</sup>, Norliza Mohd Noor<sup>2</sup>, Norulhusna Ahmad<sup>3</sup>, Taghreed Justinia<sup>4</sup>, Sahnus Usman<sup>5</sup>  
Razak Faculty of Technology and Informatics, Universiti Teknologi Malaysia  
Kuala Lumpur Campus, Jalan Sultan Yahya Petra, Malaysia<sup>1,2,3,5</sup>  
King Saud bin Abdulaziz University for Health Sciences, Jeddah, Saudi Arabia<sup>4</sup>

**Abstract**—The image super-resolution is utilized for the image transformation from low resolution to higher resolution to obtain more detailed information to identify the targets. The super-resolution has potential applications in various domains, such as medical image processing, crime investigation, remote sensing, and other image-processing application domains. The goal of the super-resolution is to obtain the image with minimal mean square error with improved perceptual quality. Therefore, this study introduces the perceptual loss minimization technique through efficient learning criteria. The proposed image reconstruction technique uses the image super-resolution generative adversarial network (ISRGAN), in which the learning of the discriminator in the ISRGAN is performed using the EfficientNet-v2 to obtain a better image quality. The proposed ISRGAN with the EfficientNet-v2 achieved a minimal loss of 0.02, 0.1, and 0.015 at the generator, discriminator, and self-supervised learning, respectively, with a batch size of 32. The minimal mean square error and mean absolute error are 0.001025 and 0.00225, and the maximal peak signal-to-noise ratio and structural similarity index measure obtained are 45.56985 and 0.9997, respectively.

**Keywords**—Single image super-resolution (SISR); generative adversarial networks (GAN); convolutional neural networks (CNN); EfficientNetv2

## I. INTRODUCTION

With the rapid development of information technology (IT) along with the boom of Internet technology, information processing based on image and signal is widely utilized by an enormous population, in which image processing is a crucial component of information processing. Here, the role of image super-resolution (SR) is significant when considering image processing-based applications [1], [2]. Image SR is the image transformation from low resolution (LR) to high resolution (HR) for obtaining an enhanced quality image. Medicine, agriculture, industry, and military applications utilize the SR technique due to its high practicability [3], [4]. While considering artificial intelligence, the role of SR is crucial for performing various processes [5], like public security, remote sensing imaging, medical imaging, image compression, and so on using the single image SR criteria [6], [7]. The image resolution enhancement using the up-sampling process lacks texture details. The image transformed into the HR provides enormous information with finer details [8]. For example, the crime scene image offers plenty of evidence for investigating crime. Likewise, an image acquired from the satellite image undergoes various processing, like resource detection, object detection, and several other processing using the HR image [9]. While considering the medical application domain, the disease diagnosis is employed based on Magnetic Resonance Imaging

(MRI) and Computed Tomography (CT) Scan images with better resolution for providing accurate medication. Thus, the role of SR is crucial in image-processing application domains.

The SR of the image is obtained from the LR image using three various categories: 1) learning-based approach, 2) reconstruction approach, and 3) interpolation approach [10]. Image resolution enhancement using interpolation is the earliest method most researchers utilized and is easy to implement. Some of the interpolation techniques used to enhance the image's quality are non-uniform sampling interpolation, Bicubic Interpolation, Bilinear Interpolation, and Nearest Neighbor Interpolation. In these approaches, the higher frequency details can be reconstructed through the linear characteristics of the approaches. The image SR using the interpolation approach provides a better outcome; still, the performance degrades with the scaling factor's elevation [11]. Reconstruction based SR approach transforms the LR image by gathering the non-redundant details. The approaches with non-negativity, energy boundedness, support boundedness, and smoothness-based hypothetical constraints using the Projection onto Convex Set are one of the methods utilized for SR reconstruction. The slower convergence rate is the limitation of the reconstruction-based criteria and has many solutions [12], [13]. In addition, the reconstruction solution acquired at the final stage hang-on on the initial evaluation. Also, the performance is limited while considering reliable robustness and real-time modeling due to the inefficiency in handling the noise level reconstruction [14]. Finally, the third approach is the learning-based image transformation approach that enhances the quality of the image using machine learning and deep learning algorithms [15]. Nowadays, learning-based image quality enhancement is the widely utilized approach by researchers [16] due to better image perception.

The learning of the network is employed in the learning-based approach of image SR for providing a high-quality reconstructed image. Here, for network learning, the high representation of the samples is utilized with variation in data for generalizing [17]. Thus, enormous data is acquired from various sources to obtain the required solution. While considering the image taken from the remote sensing domain, the information collection is a challenging task due to the variations of image based on the factors like different sensors and locations along with the difference in the objects [18], [19], [20]. Thus, the network learned with the limited samples affects the model's performance due to the poor generalization capability. Nowadays, the advent of deep learning models based on non-linear operation in artificial intelligence accumulates enormous

samples that can be utilized for several computer vision-based applications [21]. Hence, the generalization capability of the deep learning models is elevated through the enormous learning of data samples. The convolutional neural network (CNN) based image resolution enhancement technique was first suggested by Dong et al. [22]. The method acquired a better signal-to-noise ratio (SNR) with a high sampling rate, but the information about the image is lost due to the too smoothness of the reconstructed image [23], [24]. In the deep learning approaches, the error concerning the mean squared error (MSE) reduction of 2.98 and the elevation of the peak signal-to-noise ratio (PSNR) value of 21.15dB are obtained through the optimization target for acquiring the high-quality image from the ill-posed nature. The variation between the reconstructed and the original image based on the perception is not photorealistic by some existing approaches. Generative adversarial network (GAN) based methods are also designed by several researchers to overcome the issue concerning information loss in image reconstruction, in which the CNN is replaced with the GAN network [3]. The GAN-based image SR techniques gather the non-linear high-dimensional features from enormous data [25], [26] and make the generalization more effective for providing a better outcome.

This study aims to reconstruct the SR image from the LR image by minimizing the perceptual loss to obtain a perceptually good image with detailed information that gratifies the human eye. This study proposes ISRGAN with EfficientNet-v2 to enhance reconstruction accuracy with minimal distortions. The ISRGAN is designed for image reconstruction by minimizing the perceptual loss comprising of content and adversarial loss with EfficientNet-v2 is utilized for learning the discriminator of the ISRGAN, in which the perceptual loss concerning the image is minimized for the acquisition of a better quality image.

The significant contributions of the research are:

- Design of ISRGAN: Image Super-Resolution Generative Adversarial Network (ISRGAN) is designed for image reconstruction by minimizing the perceptual loss comprising of content and adversarial loss.
- EfficientNet-v2: EfficientNet-v2 is utilized for learning the discriminator of the ISRGAN, in which the perceptual loss concerning the image is minimized for the acquisition of a better quality image through the fast learning criteria.

The manuscript is organized as follows: The primary methods review is enclosed in Section II with its problem definition. The detailed proposed ISRGAN based on EfficientNet-v2 is portrayed in Section IV, Section V for implementation, and Section VI discusses the results and analysis of the study. Finally, Section VII concludes the work.

## II. RELATED WORK

The primary methods concerning the image SR are detailed in this section. The image SR using the GAN concerning the quality loss was devised by Zhu et al. [27] to solve the issue of instability of the GAN in image reconstruction. The method addressed the loss issue by optimizing the pre-trained network. The gradient magnitude similarity deviation was utilized for

the discriminator training to obtain a better performance. In addition, batch normalization, computation overhead, and memory reduction make learning more efficient and obtain better visual appeal. Image SR using GAN with the fused attentive network to extract global features was designed by Jiang et al. [28] using the scanning time reduction technique. In this study, the attributes at various scales are taken out based on the attention criteria to acquire the most relevant attributes and enhance reconstruction accuracy. In addition, the stability of the learning is accomplished through the spectral normalization approach. The analysis based on the distance measure and the similarity index offered a better outcome.

Shahidi [29] designed the wide attention-based GAN to stabilize the learning process using the Wasserstein with a Gradient penalty. Two different normalizations, such as weight and batch normalization, elevate the similarity index of the image by considering the texture and color restoration. Besides, the inclusion of the self-attention layers and the residual blocks assures a high-quality image by learning past information. The self-attention layer is utilized in the discriminator to extract the attributes from the image patches to correct the generator more effectively. The loss function was evaluated using the VGG-19 network. An enlightened GAN was developed by Gong [30] using the self-supervised hierarchical perception loss to acquire enhanced image reconstruction performance through network induced convergence. The enlighten blocks were introduced to accomplish a better gradient using the improved generalization capability. Besides, the occurrence of seam lines in the reconstructed image was eliminated through the clipping and merging approach-based learning criteria using the batch internal inconsistency loss. The image quality assessment of the method provided a superior outcome by solving the merging issue that depicts the realistic criteria of the developed method. GAN with cascading residual network was designed by hybridizing the neural network (NN) with the GAN by Ahn et al. [31] to acquire the network's short connection and multi-level representation for improving the reconstruction performance. The balanced distortion and the perception criteria of the designed method are made using the multi-scale discriminator and GAN-based learning. Here, the usage of the multi-scale discriminator gathers fine information from the image concerning the resolution to reduce the losses in the network. The visual outcome of the method provided a realistic and sharp image.

Residual-in-Residual Dense Block-based GAN was devised by Song et al. [32], in which the optimal network with transfer learning was employed to acquire a high-quality image with better perception and low distortion. The slow convergence rate and the unstable learning capability were enhanced through the inclusion of the dense block with enhanced perception and minimal distortion in the reconstructed image. In addition, the feature reuse and propagation were improved along with the vanishing gradient issue solution, which was accomplished through the dense connection establishment with minimal parameters. A conditional GAN-based image reconstruction was presented by Sun et al. [33], in which the color space transformation was initially employed based on the variance and mean of the channel for the channel normalization. The better visual representation of the image was accomplished through the colorization training criteria, which converge faster than the traditional image reconstruction approach. In addition,

curriculum learning was utilized to solve the issue of large resolution differences in the learning phase. Color normalization through self-supervised learning offers better generalization capability to obtain better performance by eliminating color variance. A weighted Multi-Scale Residual Block-based image reconstruction strategy was introduced by Li et al. [34] for image reconstruction through the weighted feature representation approach. In this, the low-level attributes were taken out from the image. Then, using the global residual learning, the high-frequency information from the features is mapped using the non-linear mapping criteria. The reconstruction of the image was accomplished from the reconstruction subnet of the network. The benefits and the challenges of the reviewed techniques is depicted in Table I.

### A. Research Gap

The acquisition of high resolution image from the low resolution input image is the image super resolution technique. The learning based methods of image super resolution are widely utilized for reconstructing the image to accomplish the better performance through the non-linear learning capability. The goal behind the image reconstruction is to obtain the better perceptual image to identify the target solution in medical imaging and satellite imaging applications. Many of the prior researchers devised the GAN based image reconstruction technique for obtaining better performance. Still, the loss function estimation and the network training with minimal loss is a challenging task in most of the methods based on the GAN. In addition, the un-pleased image reconstruction with higher distortions and the over-smoothing of the image degrades the better reconstruction. Hence in order to overcome the above mentioned challenges, an ISRGAN is introduced for learning the discriminator with higher learning speed with minimal information loss, which enhances the generalization capability of the network and obtains the better image.

## III. METHODOLOGY

The ISRGAN is designed by modifying the loss function of the traditional GAN to enhance the image quality that is more pleasant to the human eyes and to improve the accuracy of image reconstruction. The GAN for the super image resolution of the single image is termed ISRGAN for transforming the LR image into the SR image. The ISRGAN comprises of discriminator and generator, which denotes  $A_{\theta_A}$  the discriminator model. The min-max issue of the ISRGAN is solved by optimizing  $Q_{\theta_Q}$  alternatively and  $A_{\theta_A}$ . The formulation for the min-max issue is expressed as,

$$\min_{\theta_{\infty}} \max_{\theta_1} E_{S^R \sim p-\text{train}}(S^R) [\log A_{\theta_A}(S^R)] + E_{S^L \sim p-G}(S^L) | [\log (1 - A_{\theta_A}(D_{\theta_D}(S^R)))] \quad (1)$$

where, the high-resolution image is notated as  $S^R$ , the low-resolution image is indicated as  $S^L$ , the discriminator is notated as  $A$ , and the generator is indicated as  $Q$ . Here, the role of the generator is to mislead the discriminator, in which the discriminator module of the ISRGAN is learned to identify the difference between the real image and the SR image. Similarly,

TABLE I. REVIEWED TECHNIQUES: BENEFITS AND CHALLENGES

Technique	Benefits	Challenges
GAN [27]	The enhanced adversarial learning along with the loss function estimation using the image quality assessment provided a high quality image.	The failure case of the method was higher and provides an unsatisfactory result. Besides, the imperfection of the network portrayed diagonal lines in the blocks of the outcome.
Fused Attentive GAN [28]	The assessment with various GAN-based image super-resolution acquired better performance in terms of signal-to-noise ratio using various datasets. Besides, the reconstructed image is near to the real image with high resolution.	The reduction of the loss function with a higher number of iterations elevates the computation overhead.
Wide-Attention GAN [29]	Two normalizations like weights and batch normalization along with the self-attention layer provide the method a small range of gradient clipping that elevates the accuracy of image reconstruction.	The analysis without the pre-processing step limits the performance of the model.
Enlighten-GAN [30]	The convergence stabilization along with the network intensification provided a better outcome with an artifact-free reconstructed image by maintaining the hue and shape.	The outcome accomplished using the method still has the issue of an unclear outline due to the object fusing technique. Hence, for the complex image, the developed method provided poor performance.
Cascading Residual Network-GAN [31]	The improved performance is accomplished for the image translation technique by balancing distortion and perception.	The complexity analysis based on the execution time depicts the minimal time complexity compared to some baseline techniques; still, acquired poor performance due to the memory fragmentation criteria that limit the parallelism of the approach.
Residual-in-Residual Dense Block-GAN [32]	The visual outcome of the method was better and was applicable for the applications that don't require the details concerning the place and obtained better running time.	The signal-to-noise ratio evaluated by the method was not pleasant to the human eye.
Weighted Multi-Scale Residual Block [33]	The weight-based image reconstruction obtained better performance using the high-frequency attributes acquired from the low-resolution image.	The method failed to analyze the accuracy.
Conditional GAN with color normalization [34]	The augmentation of the features helps to acquire various degradations degrees, which assures the capability of learning from the de-blurred image that enhances the convergence rate and boosts the performance of the method.	For the images with severe information loss, the method offered an unsatisfactory outcome.

the generator module is trained to produce the SR image close to the real image, making the discriminator's detection capability a tedious task. As mentioned above, the generator and the discriminator processes provide a more appropriate perceptual solution. The maximization issue is solved using the discriminator with minimal MSE. The architecture of the ISRGAN is portrayed in Fig. 1.

## IV. PROPOSED WORK

The proposed ISRGAN with EfficientNet-v2 incorporates the EfficientNet-v2 to evaluate the loss function and train the discriminator based on the loss to enhance the quality of image reconstruction. The evaluation of the loss function and the

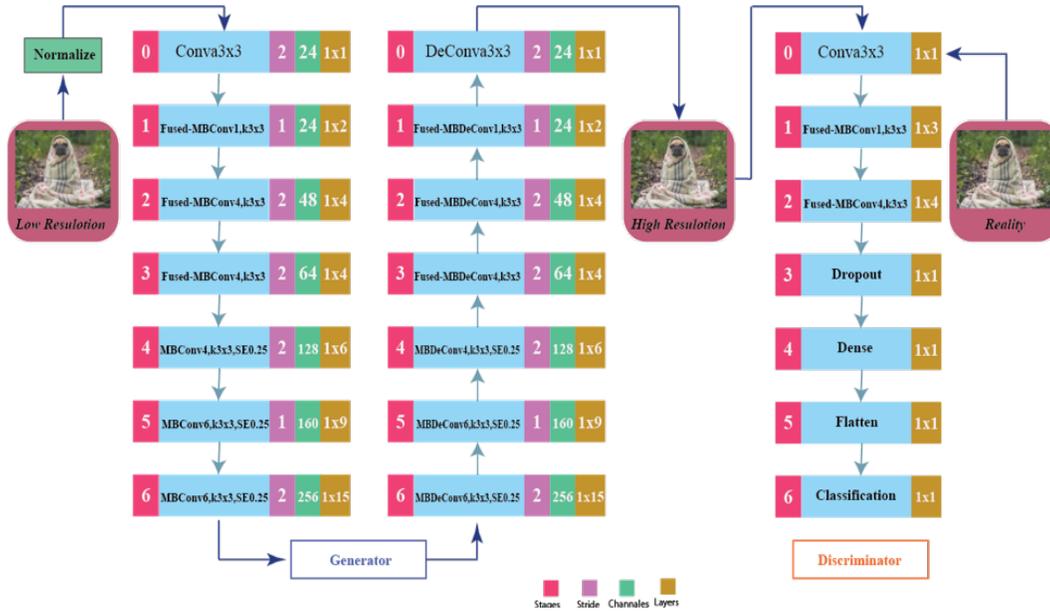


Fig. 1. The architecture of ISRGAN

learning based on the loss function using the EfficientNet-v2 is detailed below.

#### A. Loss Function

1) *Perceptual loss* : The perceptual loss function of the ISRGAN is considered a crucial factor that elevates the perceptual characteristics of the image based on the loss function [4]. The adversarial and content loss function of the ISRGAN constitutes the perceptual loss function and is formulated as,

$$h^H = h_B^H + 10^{-3}h_G^H \quad (2)$$

where, the adversarial loss of the ISRGAN is referred to as  $10^{-3}h_G^H$ , the content loss of the ISRGAN is indicated as  $h_B^H$  and the perceptual loss evaluated by the VGG-based network is indicated as  $h^H$ .

2) *Content loss*: Here, the goal of the ISRGAN is to replace the perceptual loss using the VGG feature map of the network. The loss function of the pixel-wise MSE, evaluated by VGG, is formulated as,

$$h_E^H = \frac{1}{s^2KN} \sum_{m=1}^{sK} \sum_{n=1}^{sN} (S_{m,n}^R - Q_{\theta Q}(S^L)_{m,n})^2 \quad (3)$$

where, the MSE error estimated by the VGG is indicated as  $h_E^H$ , and the height and width of the low-resolution image are indicated as  $N$  and  $K$ , respectively. For the image points  $(m, n)$ , the gray level value is indicated as  $(S^L)_{m,n}$ . Then, the characteristic loss of the VGG network is formulated as,

$$h_{VGG/u,v}^H = \frac{1}{K_{u,v}N_{u,v}} \sum_{m=1}^{K_{u,v}} \sum_{n=1}^{N_{u,v}} (\alpha_{u,v}(S^R)_{m,n} - \alpha_{u,v}(Q_{\theta Q}(S^L))_{m,n})^2 \quad (4)$$

where,  $H$  refers to the image SR,  $R$  refers to the high-resolution image,  $h_{VGG/u,v}^H$  refers to the low-resolution image,  $L$  refers to the characteristic loss for the VGG network, the reference image is indicated as  $(S^R)$ , and the reconstructed image is indicated as  $Q_{\theta Q}$ .  $u^t h$  Max-pooling layer with  $v^t h$  convolution corresponding to the feature map is indicated as  $\alpha_{u,v}$ . The dimensions of the feature maps are indicated as  $N_{u,v}$  and  $K_{u,v}$ , respectively. Here, the characteristics loss function is evaluated based on the activation function of VGG loss, which is nothing but the Euclidean distance between the reference and reconstructed image. The combination of the characteristic loss and the MSE error constitutes the content loss of the ISRGAN.

3) *Adversarial loss*: The loss generated by the generative component of the ISRGAN for providing the most appropriate solutions by tricking the discriminator of the network is termed Adversarial loss, which is formulated as,

$$h_G^H = \sum_{c=1}^C -\log A_{\theta A}(Q_{\theta Q}(S^L)) \quad (5)$$

where, the natural high-resolution image is indicated as  $Q_{\theta Q}(S^L)$ , and the reconstructed image's probability is notated as  $A_{\theta A}(Q_{\theta Q}(S^L))$ . Here, the adversarial loss function needs to be minimized for the acquisition of better gradient behavior.

### B. EfficientNet-v2 based Perceptual Loss Function

The proposed improved ISRGAN utilizes EfficientNet-v2 to estimate the loss function to enhance the quality of the reconstructed image perceptually. The ISRGAN has the issue of gradient disappearance that makes information losses due to poor generalization capability. Thus, the EfficientNet-2 is utilized for learning the discriminator based on the perceptual loss function through the higher learning rate and the parameter efficiency. Here, the role of the EfficientNet-v2 is to detect the difference between the ground truth and the reconstructed image for measuring the loss function to train the discriminator to make the reconstruction more effective. The loss functions like content loss and the adversarial losses that constitute the perceptual loss are evaluated using the EfficientNet-v2 to enhance the efficiency of the image SR.

EfficientNet-v2 for estimating the loss function utilizes smaller kernel sizes with minimal memory access overhead. Besides, the last stride-1 is eliminated to reduce the memory access overhead and size of the parameter. The runtime overhead is minimized through network capacity elevation, and the training overhead, along with higher memory, is minimized by restricting the interference of the image. The learning capability of the EfficientNet-v2 is higher by making the original interference size for learning, which depicts the scaling characteristics of the loss function estimator. The building blocks of the EfficientNet-v2 are Fused MBCConv along with the mobile inverted bottleneck MBCConv and are portrayed in Fig. 2.

### C. ISRGAN Architecture

The ISRGAN comprises two different modules: the discriminator and the generator for producing the perceptually quality image.

1) *Generator*: The transformation of the input image to obtain the SR image is employed in the generator module of the proposed ISRGAN. Initially, the encoding of the input image is performed to acquire the high-dimensional information. Then, the decoding is employed in the hidden layer of the generator to provide high-resolution images. Here, the image dimensions are scaled using the scaling algorithm to acquire the feature resolution, which is performed as convolutional and deconvolutional operations in the generator. Thus, the CNN is utilized in the generator module of the ISRGAN to extract the relevant features with a single dimension, which leads to the failure in acquiring the granular patterns of the image. Thus, the compound scaling factor is introduced in the ISRGAN to obtain the image's finer granular patterns. Hence, the generator module is designed with 42 layers of convolutional and deconvolutional, along with the Leaky rectified linear unit (Leaky ReLU) activation function and batch normalization. Besides, training aware Neural Architecture Search (NAS) and scaling are utilized for designing the generator module.

2) *Discriminator*: The difference between the high-resolution image generated by the generator module and the real image is identified by the discriminator module using the EfficientNet-v2. Here also, finer granular patterns are extracted from the image. Seven Fused-MBCConv, one convolutional layer, and eight hidden layers constitute the discriminator module along with the EfficientNet-v2. The sigmoid function

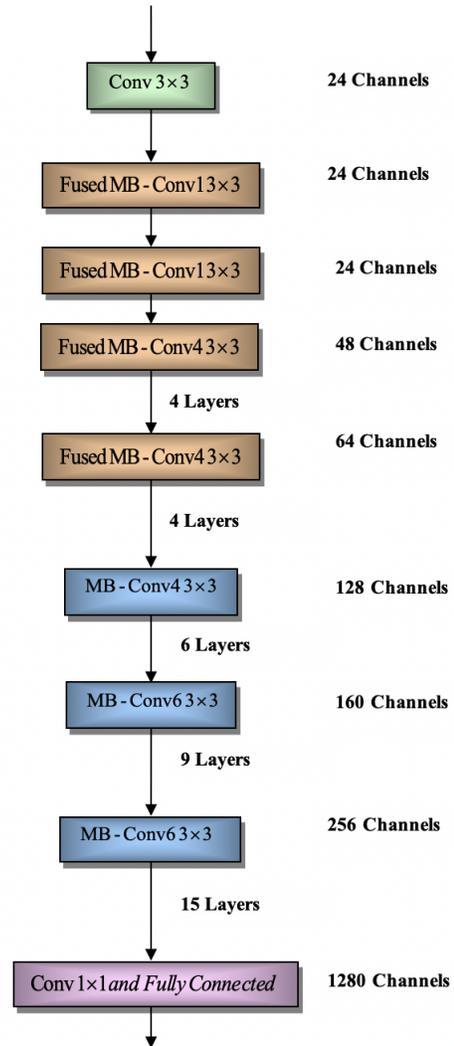


Fig. 2. Architecture of efficientNetv2

is utilized to determine the likelihood of SR images by considering the fake and real images. The perceptual loss function training is employed alternatively in both the discriminator and generator to obtain perceptually better image SR.

## V. IMPLEMENTATION

The experimentation setup along with the interpretation of the introduced ISRGAN is detailed in this section.

### A. Experimental Setup

The implementation of the proposed method is employed in Google Colaboratory (Colab) Notebooks tool. The total number of iterations performed is 100, with batch sizes of 8 and 32.

1) *Image dataset*: The dataset utilized for the analysis of the proposed ISRGAN is the CIFAR-10 dataset [35] comprises 10000 test images and 50,000 training images with 10 different classes. Thus, a total of 60,000 images with the size of is utilized. The type of image utilized for the evaluation is synthesis.

2) *Network learning*: The generator module of the ISRGAN is learned with the loss function, expressed as,

$$\text{Loss}_Q = h^H + \text{Loss}_{\text{texture}} \quad (6)$$

where,  $\text{Loss}_Q$  refers to the loss utilized to learn the generator,  $h^H$  notates the perceptual loss, which comprises adversarial loss and content loss and the texture loss is indicated as  $\text{Loss}_{\text{texture}}$ . Here, the Adam optimizer is utilized to solve the optimization issue, and the learning rate of  $1e^{-4}$  is initially utilized. EfficientNet-v2 is utilized for extracting the finer granular patterns from the image.

3) *Loss function*: The empirical risk of the proposed ISRGAN in reconstructing the HR image is evaluated based on the loss function and is enunciated as,

$$\text{Loss} = \frac{1}{W} \sum_{i=1}^W (O_i - \hat{O}_i)^2 \quad (7)$$

where, the total reconstructed samples is indicated as  $W$ , the HR reconstructed image by the proposed ISRGAN is  $O_i$  and the targeted outcome is indicated as  $\hat{O}_i$ .

4) *Network parameter setting*: The setting for the network parameters of the Generator and Discriminator in the ISRGAN architecture are displayed in Table II.

TABLE II. ISRGAN PARAMETER SETTINGS

Parameters	Values
Generator	
Dense Layer	256x4x4
Conv-1	128 filters, 3x3 kernel, 2x2 stride
Conv-2	128 filters, 3x3 kernel, 2x2 stride
Conv-3	128 filters, 3x3 kernel, 2x2 stride
Conv-4	3 filters, 3x3 kernel
Discriminator	
Input Layer	32x32x3
Conv-2	64 filters, 3x3 kernel, 2x2 stride
Conv-3	128 filters, 3x3 kernel, 2x2 stride
Conv-4	128 filters, 3x3 kernel, 2x2 stride
Conv-5	256 filters, 3x3 kernel, 2x2 stride

## VI. RESULTS AND ANALYSIS

### A. Qualitative Analysis

The experimental outcome concerning the input LR image and the output HR image is shown in Fig. 3.

The outcome acquired by the introduced ISRGAN shown in Fig. 3(b) is visually pleasing to the human eye by reconstructing the LR image 3(a). The discriminator learning based on the perception loss estimated by the generator using the EfficientNet-v2 elevates the accuracy of image reconstruction by minimizing the MSE.

### B. Quantitative Analysis

The quantitative analysis based on the loss function evaluated by the proposed method and the error analysis is detailed in this section.



Fig. 3. Outcome of ISRGAN: (a) input LR image and (b) output HR image

1) *Analysis based on loss function*: The quantitative analysis of the introduced ISRGAN is analyzed based on the loss function measure and is portrayed in Fig. 4. The Table III presents the batch sizes, number of iterations, and outcomes of generator, discriminator, and the self-supervised. From the table, we can observe that the increase in batch size and iteration minimizes the loss. Here, the increase in the batch size of the method reduces the noise in the gradients, which assures enhanced accuracy in image reconstruction and reduces the loss. Similarly, the increase in iteration helps the algorithm obtain a closer solution to the target solution, which reduces the loss and enhances image reconstruction accuracy. The ISRGAN with the EfficientNet-v2 acquires minimal loss and can be applied for the media, medical, and surveillance-related application domains by detecting various objects in the image. Besides, the loss function evaluation based on the perceptual similarity loss assists in capturing the high-level information that provides a more accurate solution.

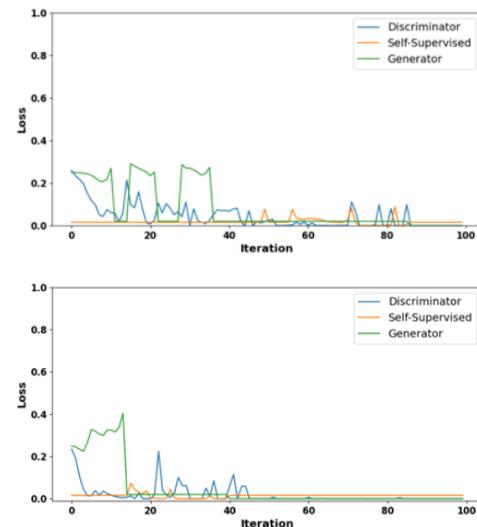


Fig. 4. Analysis based on loss function with batch size 8 (top) and batch size 32 (bottom)

TABLE III. LOSS FUNCTION ANALYSIS

Batch size	Iteration	Generator	Discriminator	Self supervised
8	20	0.251	0.012	0.0154
32	20	0.02	0.000057	0.037
8	100	3.778	3.916	0.0154
32	100	0	9.130	0.015

2) *Error and similarity analysis:* The analysis is executed based on MSE, PSNR, structural similarity index measure (SSIM), and mean absolute error (MAE).

**MSE:** The analysis of the risk function concerning the target solution in reconstructing the image is measured through the MSE and is formulated as,

$$ISRGAN_{MSE} = \frac{1}{T_{samples}} \sum_{i=1}^{T_{samples}} (I_{outcome} - I_{Target})^2 \quad (8)$$

Here, the targeted outcome is indicated as  $I_{Target}$ , the outcome acquired by the proposed ISRGAN is indicated as  $I_{outcome}$ , the total number of samples is indicated as  $T_{samples}$  and the MSE is notated as  $ISRGAN_{MSE}$ . The MSE evaluated by the proposed ISRGAN and the comparative methods like Wide-Attention GAN [29], Enlighten GAN [30], and Fused Attentive GAN [28] is portrayed in Fig. 5. The MSE evaluated by the introduced ISRGAN with the EfficientNet-v2 is minimal compared to the other conventional methods.

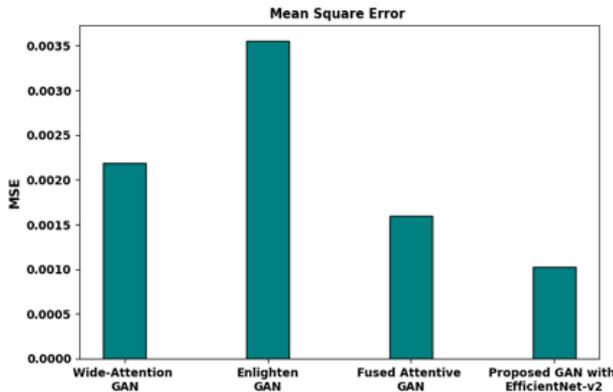


Fig. 5. Analysis based on MSE

**MAE:** The absolute difference between the obtained and the targeted outcome of the super image resolution is measured through the MAE and is formulated as,

$$ISRGAN_{MAE} = \frac{1}{T_{samples}} \sum_{i=1}^{T_{samples}} |I_{outcome} - I_{Target}| \quad (9)$$

where, the MAE measure is indicated as  $ISRGAN_{MAE}$ . The outcome of the proposed ISRGAN, along with the comparative methods, is portrayed in Fig. 6. The MAE evaluated by the proposed method is minimal compared to the other state-of-the-art techniques.

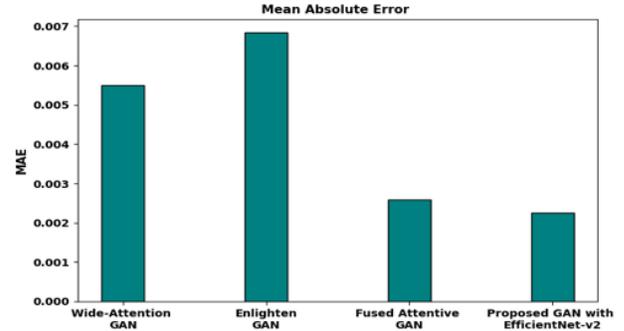


Fig. 6. Analysis based on MAE

**PSNR:** The quality of the image reconstructed from the low-resolution image is measured through the PSNR. The greater value of PSNR depicts better reconstruction. The formulation of the PSNR of the reconstructed image is expressed as,

$$ISRGAN_{PSNR} = 10 \log_{10} \left( \frac{R^2}{ISRGAN_{MSE}} \right) \quad (10)$$

where, the PSNR measure is indicated as  $ISRGAN_{PSNR}$ , and the fluctuation of the reconstructed image compared to the original image is denoted as  $R$ . The PSNR estimated by the proposed and the comparative methods is portrayed in Fig. 7. The PSNR acquired by the introduced method is higher and depicts a better image reconstruction quality.

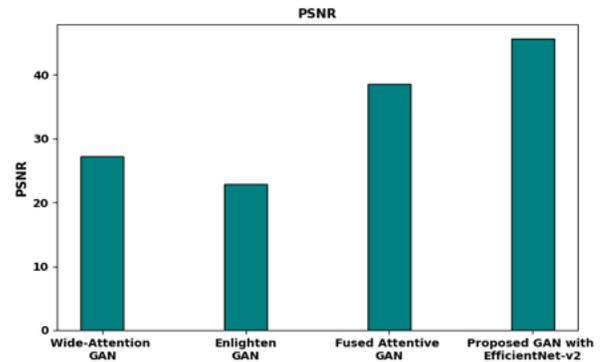


Fig. 7. Analysis based on PSNR

**SSIM:** The similarity between the reconstructed and the original image is measured through the SSIM that depicts the perceptual quality of the image. The mathematical formulation of the SSIM is expressed as,

$$ISRGAN_{SSIM} = \frac{(2\mu_p\mu_q + s_1)(2\sigma_{pq} + s_2)}{(\mu_p^2 + \mu_q^2 + s_1)(\sigma_p^2 + \sigma_q^2 + s_2)} \quad (11)$$

where, the SSIM is indicated as  $ISRGAN_{SSIM}$ . The mean and the variance of the reconstructed image are indicated

as  $\mu_p$  and  $\sigma_p^2$ , and the mean and variance of the original image are indicated as  $\mu_q$  and  $\sigma_q^2$ . The covariance between the original and reconstructed image is notated as  $\sigma_{pq}$ , and the variable utilized for the stabilization is indicated as  $s_1$  and  $s_2$ , respectively.

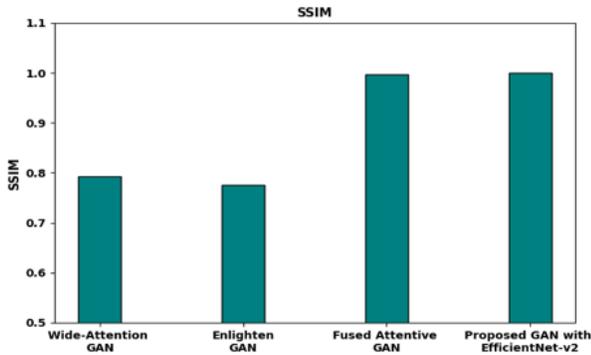


Fig. 8. Analysis based on SSIM

The outcome of the proposed method, along with its comparative methods, is portrayed in Fig. 8. The higher SSIM evaluated by the proposed method depicts the similarity of the reconstructed super resolution image with the original image.

3) *Comparative discussion:* The comparative analysis of the proposed method, along with the conventional image reconstruction techniques, is portrayed in Table IV. The MSE acquired by the proposed method is 0.001025, which is 52.93%, 70.99%, and 35.54% minimum compared to the traditional Wide-Attention GAN, Enlighten GAN, and Fused Attentive GAN methods. Similarly, the proposed method accomplished the minimal MAE of 0.00259, which is 59.09%, 67.08%, and 13.13% minimum compared to the traditional Wide-Attention GAN, Enlighten GAN, and Fused Attentive GAN methods. In contrast, the proposed method acquired the maximal PSNR of 45.5698, which is 40.26%, 49.89%, and 15.34% better compared to conventional Wide-Attention GAN, Enlighten GAN, and Fused Attentive GAN methods. Likewise, the maximal SSIM measured by the proposed method is 0.9997, which is 20.62%, 22.32%, and 0.18% superior compared to state-of-the-art Wide-Attention GAN [29], Enlighten GAN [30], and Fused Attentive GAN [28] methods.

TABLE IV. COMPARATIVE ANALYSIS

Method	MSE	PSNR	SSIM	MAE
Wide-Attention GAN [29]	0.002188	27.2254	0.7936	0.0055
Enlighten-GAN [30]	0.00355	22.834	0.77656	0.006835
Fused Attentive GAN [28]	0.001598	38.58	0.9979	0.00259
Proposed Method	0.001025	45.56985	0.9997	0.00225

The proposed method incorporates the EfficientNet-v2 with the traditional GAN for enhancing the image super resolution quality by minimizing the perceptual loss. For this, the EfficientNet-v2 is utilized at the discriminator that minimizes the information loss with fast learning rate. Thus

The Fused GAN was devised by integrating the attention module in the traditional GAN for enhancing the performance of the image super resolution technique. The quality assessment based on the similarity measure was closer to the proposed method; still, the higher time complexity and failure in maintaining the system parallelism mechanism limits the method. Thus, the existing Fused GAN acquired little minimal performance compared to the proposed GAN with EfficientNet-v2. Followed by, Wide Attention GAN devised by integrating the self-attention module. Due to the inclusion of the attention mechanism, the introduced Wide Attention GAN also suffered with higher time complexity. Finally, the Enlighten-GAN that was designed by incorporating enlighten blocks into the traditional GAN. Here, the introduced method accomplished better quality image as outcome by solving the unstable convergence issue through the enlighten block. However, method is inefficient in processing the image with complex background. Thus, the analysis based on the error estimation and the similarity analysis shows that the proposed method accomplished better performance than other super image resolution methods. The training of the discriminator using the EfficientNet-v2 provides fast learning and the least loss function based on the perception enhancing the image's visual quality.

## VII. CONCLUSION

The SR image reconstruction from the LR image for the acquisition of fine-grained information from the image, along with the visual quality, is attained using ISRGAN with EfficientNet-v2 in this paper. The proposed image reconstruction technique elevates the quality of the image by minimizing the perception loss present in the GAN network. The information loss is minimized through the efficient learning of the discriminator using the EfficientNet-v2, which has the probability of a fast learning rate with more accurate learning based on the perception loss generated by the generator module of the ISRGAN. The learning capability of the EfficientNet-v2 is higher by making the original interference size for learning that depicts the scaling characteristics of the loss function estimator. Thus, the proposed method provides a promising outcome based on the error analysis and acquires the minimal values of loss of 0.02 at the generator, 0.1 at the discriminator, and 0.015 at the self-supervised learning, respectively, with a batch size of 32. Besides, the minimal MSE and MAE accomplished by the proposed method are 0.001025 and 0.00225, respectively. Likewise, the maximal PSNR and SSIM acquired by the proposed method are 45.56985 and 0.9997, respectively. However, the loss of the method further needs to be reduced for real-time processing.

In the future, we plan to introduce a novel hybrid meta-heuristic optimization-based deep learning approach to overcome the challenges the proposed model encountered in this study.

## REFERENCES

- [1] J. Xiu, X. Qu, and H. Yu, "Face super-resolution using recurrent generative adversarial network," in *2022 IEEE 6th Information Technology and Mechatronics Engineering Conference (ITOEC)*, vol. 6. IEEE, 2022, pp. 1169–1174.

- [2] X. Lin, X. Zhou, T. Tong, X. Nie, L. Wang, H. Zheng, J. Li, E. Xue, S. Chen, M. Zheng *et al.*, "A super-resolution guided network for improving automated thyroid nodule segmentation," *Computer Methods and Programs in Biomedicine*, p. 107186, 2022.
- [3] A. Abedjooy and M. Ebrahimi, "Multi-modality image super-resolution using generative adversarial networks," *arXiv preprint arXiv:2206.09193*, 2022.
- [4] S. Altakrouri, S. B. Usman, N. B. Ahmad, T. Justinia, and N. M. Noor, "Image to image translation networks using perceptual adversarial loss function," in *2021 IEEE International Conference on Signal and Image Processing Applications (ICSIPA)*. IEEE, 2021, pp. 89–94.
- [5] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, "Photo-realistic single image super-resolution using a generative adversarial network," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 105–114.
- [6] C. Saharia, J. Ho, W. Chan, T. Salimans, D. J. Fleet, and M. Norouzi, "Image super-resolution via iterative refinement," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.
- [7] Y. Li, B. Sixou, and F. Peyrin, "A review of the deep learning methods for medical images super resolution problems," *IRBM*, vol. 42, no. 2, pp. 120–133, 2021. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1959031820301408>
- [8] B. Meng, L. Wang, Z. He, G. Jeon, Q. Dou, and X. Yang, "Gradient information distillation network for real-time single-image super-resolution," *J. Real-Time Image Process.*, vol. 18, no. 2, p. 333–344, apr 2021. [Online]. Available: <https://doi.org/10.1007/s11554-021-01083-1>
- [9] J. Hou, Y. Si, and X. Yu, "A novel and effective image super-resolution reconstruction technique via fast global and local residual learning model," *Applied Sciences*, vol. 10, no. 5, 2020. [Online]. Available: <https://www.mdpi.com/2076-3417/10/5/1856>
- [10] Y. Sun, R. Wang, D. Cao, and R. Lee, "Who are social media influencers for luxury fashion consumption of the chinese gen z? categorisation and empirical examination," *Journal of Fashion Marketing and Management: An International Journal*, 2021.
- [11] X. Liu, J. Li, T. Duan, J. Li, and Y. Wang, "Dsma: Reference-based image super-resolution method based on dual-view supervised learning and multi-attention mechanism," *IEEE Access*, 2022.
- [12] Z. Lu and Y. Chen, "Single image super-resolution based on a modified u-net with mixed gradient loss," *signal, image and video processing*, vol. 16, no. 5, pp. 1143–1151, 2022.
- [13] P. Andreini, G. Ciano, S. Bonechi, C. Graziani, V. Lachi, A. Mecocci, A. Sodi, F. Scarselli, and M. Bianchini, "A two-stage gan for high-resolution retinal image generation and segmentation," *Electronics*, vol. 11, no. 1, 2022. [Online]. Available: <https://www.mdpi.com/2079-9292/11/1/60>
- [14] K. M. Peddecord, "Behold the future: Implications of automated image analysis for the photographic arts," *The International Journal of the Image*, vol. 13, no. 1, p. 47–62, 2022.
- [15] K. Varma, G. S. Reddy, and N. Subramanyam, "Face image super resolution using a generative adversarial network," in *2021 Smart Technologies, Communication and Robotics (STCR)*. IEEE, 2021, pp. 1–8.
- [16] T. Shinohara, R. Ito, Y. Kobayashi, T. Satoh, Y. Shimazaki, and S. Nakamura, "Application of pre-trained real-world super resolution models to optical satellite image," *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 43, pp. 167–173, 2022.
- [17] T.-A. Song, S. R. Chowdhury, F. Yang, and J. Dutta, "Pet image super-resolution using generative adversarial networks," *Neural Networks*, vol. 125, p. 83–91, 2020.
- [18] Y. Tao, S. J. Conway, J.-P. Muller, A. R. Putri, N. Thomas, and G. Cremonese, "Single image super-resolution restoration of tgo cassis colour images: Demonstration with perseverance rover landing site and mars science targets," *Remote sensing*, vol. 13, no. 9, p. 1777, 2021.
- [19] Y. Kim and D. Son, "Noise conditional flow model for learning the super-resolution space," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 424–432.
- [20] H. Y. Park, H.-J. Bae, G.-S. Hong, M. Kim, J. Yun, S. Park, W. J. Chung, and N. Kim, "Realistic high-resolution body computed tomography image synthesis by using progressive growing generative adversarial network: visual turing test," *JMIR Medical Informatics*, vol. 9, no. 3, p. e23328, 2021.
- [21] M. Tan and Q. Le, "Efficientnetv2: Smaller models and faster training," in *International Conference on Machine Learning*. PMLR, 2021, pp. 10096–10106.
- [22] C. Dong, C. C. Loy, K. He, and X. Tang, "Learning a deep convolutional network for image super-resolution," in *European conference on computer vision*. Springer, 2014, pp. 184–199.
- [23] B. Liu and J. Chen, "A super resolution algorithm based on attention mechanism and srgan network," *IEEE Access*, vol. 9, pp. 139138–139145, 2021.
- [24] S. Kim, D. Jun, B.-G. Kim, H. Lee, and E. Rhee, "Single image super-resolution method using cnn-based lightweight neural networks," *Applied Sciences*, vol. 11, no. 3, p. 1092, 2021.
- [25] Y. Xiong, S. Guo, J. Chen, X. Deng, L. Sun, X. Zheng, and W. Xu, "Improved srgan for remote sensing image super-resolution across locations and sensors," *Remote Sensing*, vol. 12, no. 8, p. 1263, 2020.
- [26] X. Liu, S. Chen, L. Song, M. Woźniak, and S. Liu, "Self-attention negative feedback network for real-time image super-resolution," *Journal of King Saud University-Computer and Information Sciences*, vol. 34, no. 8, pp. 6179–6186, 2022.
- [27] X. Zhu, L. Zhang, L. Zhang, X. Liu, Y. Shen, and S. Zhao, "Gan-based image super-resolution with a novel quality loss," *Mathematical Problems in Engineering*, vol. 2020, 2020.
- [28] M. Jiang, M. Zhi, L. Wei, X. Yang, J. Zhang, Y. Li, P. Wang, J. Huang, and G. Yang, "Fa-gan: Fused attentive generative adversarial networks for mri image super-resolution," *Computerized Medical Imaging and Graphics*, vol. 92, p. 101969, 2021.
- [29] F. Shahidi, "Breast cancer histopathology image super-resolution using wide-attention gan with improved wasserstein gradient penalty and perceptual loss," *IEEE Access*, vol. 9, pp. 32795–32809, 2021.
- [30] Y. Gong, P. Liao, X. Zhang, L. Zhang, G. Chen, K. Zhu, X. Tan, and Z. Lv, "Enlighten-gan for super resolution reconstruction in mid-resolution remote sensing images," *Remote Sensing*, vol. 13, no. 6, p. 1104, 2021.
- [31] N. Ahn, B. Kang, and K.-A. Sohn, "Efficient deep neural network for photo-realistic image super-resolution," *Pattern Recognition*, vol. 127, p. 108649, 2022.
- [32] H. Song, M. Wang, L. Zhang, Y. Li, Z. Jiang, and G. Yin, "S2rgan: sonar-image super-resolution based on generative adversarial network," *The Visual Computer*, vol. 37, no. 8, pp. 2285–2299, 2021.
- [33] L. Sun, Z. Liu, X. Sun, L. Liu, R. Lan, and X. Luo, "Lightweight image super-resolution via weighted multi-scale residual network," *IEEE/CAA Journal of Automatica Sinica*, vol. 8, no. 7, pp. 1271–1280, 2021.
- [34] B. Li, A. Keikhosravi, A. G. Loeffler, and K. W. Eliceiri, "Single image super-resolution for whole slide image using convolutional neural networks and self-supervised color normalization," *Medical Image Analysis*, vol. 68, p. 101938, 2021.
- [35] "The cifar-10 dataset," (Date last accessed 20-November-2014). [Online]. Available: <https://www.cs.toronto.edu/~kriz/cifar.html>

# A Transformer Seq2Seq Model with Fast Fourier Transform Layers for Rephrasing and Simplifying Complex Arabic Text

Abdullah Alshanqiti<sup>1</sup>, Ahmad Alkhodre<sup>2</sup>, Abdallah Namoun<sup>3</sup>, Sami Albouq<sup>4</sup>,  
Emad Nabil<sup>5</sup>

Faculty of Computer and Information Systems, Islamic University of Madinah  
Madinah 42351, Saudi Arabia<sup>1,2,3,4</sup>  
Faculty of Computers and Artificial Intelligence  
Cairo University, Giza 12613, Egypt<sup>5</sup>

**Abstract**—Text simplification is a fundamental unsolved problem for Natural Language Understanding (NLU) models, which is deemed a hard-to-solve task. Recently, this hard task has aimed to simplify texts with complex linguistic structures and improve their readability, not only for human readers but also for boosting the performance of many natural language processing (NLP) applications. Towards tackling this hard task for the low-resource Arabic NLP, this paper presents a text split-and-rephrase strategy for simplifying complex texts, which depends principally on a sequence-to-sequence Transformer-based architecture (which we call TSimAr). For evaluation, we created a new benchmarking corpus for Arabic text simplification (so-called ATSC) containing 500 articles besides their corresponding simplifications. Through our automatic and manual analyses, experimental results report that our TSimAr evidently outperforms all the publicly accessible state-of-the-art text-to-text generation models for the Arabic language as it achieved the best score on SARI, BLEU, and METEOR metrics of about 0.73, 0.65, and 0.68, respectively.

**Keywords**—Text simplification; sequence-to-sequence; split-and-rephrase; natural language understanding; NLP; TSimAr; ATSC

## I. INTRODUCTION

Texts with complex linguistic structures often pose a difficulty in interpreting and understanding the indented meanings, particularly the meanings between the lines. Such difficulty is not only encountered by human readers but also by intelligent applications that demand text comprehension at some point. Consequently, text simplification methodologies have come to help various readers (mainly readers with low-literacy skills [1], such as children or non-native readers) as well as to boost the performance of many natural language processing (NLP) applications (e.g., automated text parsing [2], summarizations [3], and translations).

Given a linguistically complex text as input, automated text simplifications (ATS) work around generating candidate texts as output that are essentially uncomplicated in structure and easy to understand without losing the purposed meaning [4]. In broad, the relatively common two steps involved in many ATS approaches are 1) splitting complex texts into simple sentences [5], [6] and 2) text rephrasing [7], [8], [9], [10], [11] using more straightforward common words (well-known in some approaches as lexical paraphrasing). In this paper, we

focus on these two steps and introduce a novel text split-and-rephrase solution for Arabic ATS.

Unlike the highly supported Indo-European languages (such as English), the recent ATS literature [4], [12], [13], [14] indicates that quite a few works are dedicated to supporting the Arabic language. To our knowledge, a few text splitting (e.g. [5]) and/or text-to-text rephrasing (e.g. [15], [16], [17], [18], [19]) techniques exist in Arabic NLP literature for tasks not related to simplification problems. Thus, in this paper, we seek to combine these techniques to support Arabic ATS in general. To illustrate the originality of our proposal, in the next section, we review in some detail these techniques besides exploring the existing non-Arabic split-and-rephrase models [7], [20], [10], [9]. Until then, we briefly outline the originality and main contributions of this paper as follows:

- We introduce a text split-and-rephrase strategy for simplifying complex Arabic texts, which depends principally on a sequence-to-sequence Transformer-based architecture. In the splitting part, we integrate our suggested solution with a punctuation detector for text segmentation (PDTs) built on top of a pre-trained multilingual masked-language model (mBERT). This PDTs attempts to generate the shortest set of simple independent-clause sentences from a given lengthy complex text. While in the rephrasing part, we propose a modified attention-free Transformer model, depending on a fast Fourier-Transform (FNet-based), which rephrases the concatenated simple sentences into a more readable version. The significant (original) work introduced in this paper focuses on the latter part.
- We create a new Arabic corpus for benchmarking text simplification approaches and make it publicly available<sup>1</sup>. Besides, we make the details of our experimental evaluations and implementations (i.e., including codes and scripts) publicly accessible<sup>2</sup> to the interested researchers for replicating our experiments.

The remainder of this paper structurally unfolds as follows. First, we present the closely related work and the state-of-the-art NLP pre-trained models, including a brief review of gaps in the literature concerning the Arabic text simplification prob-

<sup>1</sup><https://github.com/AMahfodh/ArSummarizer/tree/main/TSimAr/resources>

<sup>2</sup><https://github.com/AMahfodh/ArSummarizer/tree/main/TSimAr>

lem. Next, we introduce our text split-and-rephrase approach and discuss the subsequent experimental analysis and results. Concluding the paper and outlining the potential forthcoming avenues of investigation are presented afterward.

## II. REVIEW OF RELATED LITERATURE

This section provides a synopsis of Text Simplification (TS) methods, broadly categorized into extractive and abstractive approaches. In principle, text simplification approaches (illustrated in Figure 1) are abstractly analogous to text summarization approaches, as they both strive to simplify texts. However, the dissimilarity between them is fundamental. Text summarization attempts to shorten text without losing the key meanings, while simplification attempts to improve text readability by reducing linguistic complexity constrained by preserving the key meanings. Before delving into TS approaches, we position the novelty of our contribution in the context of related literature by stating the following:

- no split-and-rephrase model has been suggested yet for simplifying complex Arabic texts to the extent of the authors' knowledge; and
- for text-to-text rephrasing, we explore, for the first time, the effect of the modified attention-free Transformer model [21](i.e., depends on a fast Fourier-Transform) for the Arabic language.

### A. Extractive Approaches

The core idea of the extractive approach is to extract the main sentences of the text through text summarization. In other words, the simplification here is performed using summarization. However, summarization does not necessarily lead to simplification. Therefore, this approach is not recommended. One of the examples of text simplification through summarization is the TF-IDF [14]. Preprocessing is a normal requirement of this approach. The preprocessing includes converting text into lower-case, removing punctuation, special characters, and stop words, and stemming to return complex words to their language base.

### B. Abstractive Approaches

The previous approach is not a real simplification, it is just a summarization, but in the abstractive approach, the output text is a real simplified text. In this approach, there are two main categories of simplifications. The first is Lexical Simplification (LS), and the second is Text Generation (TG).

1) *Lexical simplification*: The simplification in the LS category is performed through the replacement of complex/hard words with simple/easy words. Therefore, it is called lexical simplification. LS algorithms work at the sentence level. The structure of sentences is not changed, and the grammar is simplified. Only word replacement is included in the simplification. As so, this type of simplification is not effective enough. The hard words may be replaced with easy ones, but the sentence structure and grammar may still be hard to understand. Examples of lexical simplification are the following.

- Rules-based LS
- Parallel corpora extracted-rules LS
- Word embedding LS

- Pre-trained language models LS

Rule-based LS [22], [23] depends on a linguistic database, like WordNet, to get the simplest synonym of a word, which can be based on its frequency or its length. In parallel corpora, extracted-rules TS [24], [25], [26], the rules are extracted automatically from a parallel aligned corpus. While Word Embedding TS [27], [28], [29] has the advantage that there is no need for lexical resources. The appearance of pre-trained models had a huge benefit for all NLP tasks; one of these tasks is text simplification. Some systems use pre-trained models like BERT to find and generate easy words for complex ones. Therefore, the class of pre-trained language models TS systems [30], [31] has proved its effectiveness compared to other techniques.

2) *Text generation*: The second category of the abstractive approach is Text Generation (TG). In TG, a new simplified text is generated. The new text may have a different structure or a different number of sentences. TG's approach includes sentence splitting, text addition, and deletion. TG can only be considered a true simplification, while the previous methods can be seen as good trials for the simplification, but not real simplification because real simplification means digestion of the original text and generating a new simplified one with new simple words, structures, and grammar. The recent text generation-based TS is data-driven, considering the advantage of the complex structures in data. The text generation techniques can be classified as follows:

- Syntactic simplification
- Statistical Machine Translation (SMT)
- Deep Learning Techniques

In the next paragraphs, we explain each approach.

a) *Syntactic simplification*: In syntactic simplification, the hard/complex words are replaced by easy/simple ones, and the grammatically complex sentences are identified and rewritten in simple sentences. The process of simplification includes splitting long sentences, changing passive sentences to active ones, and resolving ambiguities. Examples of syntactic simplification research are mentioned in [32], [33].

In [34], the authors proposed a model called: TriS: Their approach was to break down a long statement into numerous shorter ones. When a sentence is written in the subject-verb-object order, it is called simple (SVO). A dataset of 854 sentences taken from the New York Times and Wikipedia, manual simplification is performed for the evaluation. The authors assess 100 unseen sentences and compare them to Heilman and Smith's rule-based method [35]. Their approach gets higher ROUGE [36] and Flesch-Kincaid Grade Level scores. Another method is grammar induction, where simplification is considered by converting the tree1 to tree2 problem, where tree1 is for the source, and tree2 is for the target. The process includes extracting tree transformation rules using a corpus, then learning how to select the adequate rule(s) to be applied to simplify unseen sentences. In [37], authors modeled the syntactic and lexical simplification using tree transduction rules. The evaluation of their proposal was performed using the Simple Wikipedia corpus and it showed good results. The authors highlighted the need for a mechanism to eliminate useless transformation rules.

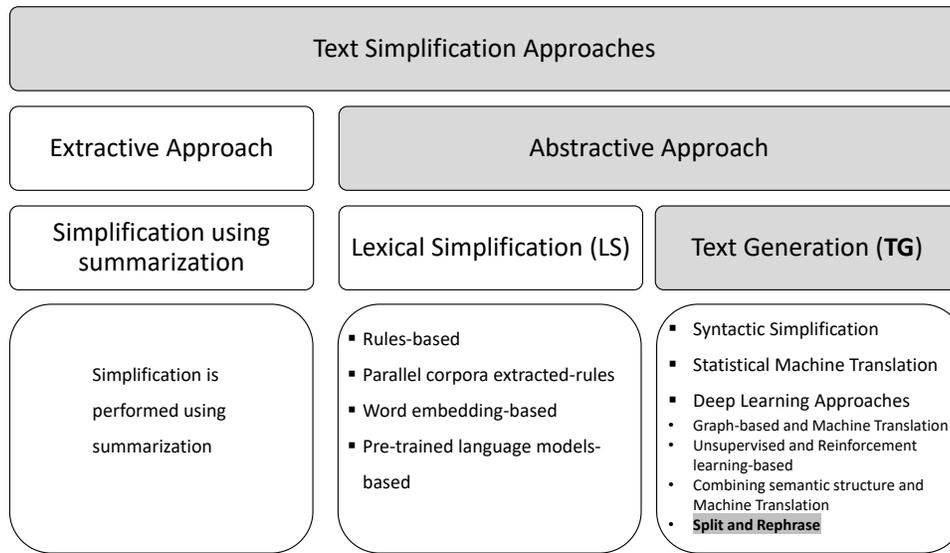


Fig. 1. Text simplification approaches

b) Statistical machine translation (SMT): Machine translation (MT) is the process of translating a text written in language A (source) to language B (target), where the two languages are different. Due to the massive amount of available data nowadays, MT has achieved many success stories. MT is applied successfully to the TS by considering that language B (target) is a simplified version of language A (source), where both represent the same language. Now the problem of simplification can be seen as the generation of monolingual text-to-text or monolingual translation. Some research used phrase-based statistical MT and applied it to TS. The task obviously is simpler than the source, and the target languages are the same; only the target is simpler than the source. Examples of SMT used for TS includes [38], [39], [40], [41], [42].

c) Deep learning techniques: In the era of big data, powerful computers, and GPUs, deep learning took the lead in AI, especially data-driven AI. Deep learning proved to be effective when used with SMT, where RNN Encoder-Decoder is used in MT [43]. This motivated researchers to employ DL in TS using the monolingual translation approach. In [44], the researchers successfully used RNN-based Neural Machine Translation (NMT) for TS; other authors in [45] used LSTM Encoder-Decoder in the simplification process. The authors in [46] developed a model called R-PBMT, Phrase-Based Machine Translation, augmented with a re-Ranking heuristic based on dissimilarity. The model is trained and tested using the PWKP dataset; they compare their work with three models, Word-Substitution Baseline models, that replace a word in a sentence with synonyms retrieved from WordNet. In another research [47], the authors performed four rewriting operations, replacing, splitting, reordering, and deletion; their work is also depending on DL. In [48], the authors proposed a system that is based on quasi-synchronous grammar. Results showed the general superiority of their model using the human evaluation and the automatic evaluation using the metrics BLEU and Flesch-Kincaid grade level. There exist many other DL-based approaches for TS. These approaches include graph-based

approaches [49], reinforcement learning-based TS [50], NMT [51], combining semantic structure and NMT [52], phrase-based unsupervised TS [53], unsupervised neural TS [54], and split-and-rephrasing techniques [8], [9], [7], [10], [20]. In this paper, we conceptually consider the latter technique (i.e., split-and-rephrasing technique) for the Arabic language for the first time. In addition, Table I summarizes the recent existing approaches which indicate the originality of our proposal.

### III. METHODOLOGY

Figure 2 presents an overview of our proposed TSimAr in five steps with illustrative input-output examples. In principle, we integrate a punctuation detector for text segmentation (PDTs [5], see step 3) with a modified attention-free Transformer architecture for rephrasing and simplifying complex Arabic text, step 4. The former (i.e., PDTs) attempts to split a given input text into the shortest set of simple independent-clause sentences. The latter (i.e. the focus of this paper) aims to rephrase the concatenated simple sentences to generate a more readable version. For example, given a textual document  $X$  containing complex sentences, TSimAr attempts to break it down into ( $Y$ ) uncomplicated sentences with rephrasing, such that  $Y \leftarrow TSimAr(X)$  and  $Y = \{y_1, y_2, \dots, y_{|n|}\}$ , where  $n$  is the number of generated simple sentences.

As usual with most NLP applications, TSimAr is set off with straightforward text preprocessing (see step 2), which includes noise/diacritizations removal and soft normalization. This preprocessing step cleans the input texts without breaking sentence structures, and more importantly, it preserves the overall meaning to an extent. For implementing this step, we consider two Python-based toolkits: NLTK<sup>3</sup> and CAMEL<sup>4</sup>. Once the preprocessing step is performed, TSimAr segments and then rephrases the cleaned input text, shown in steps 3 and 4 of Figure 2. In the following subsections, we focus on these

<sup>3</sup>NLTK Tool: <https://www.nltk.org/>

<sup>4</sup>CAMEL is an Arabic natural language processing tool: <https://camel-tools.readthedocs.io/>

TABLE I. A BRIEF SUMMARY OF REVIEWED TEXT SIMPLIFICATION APPROACHES THAT APPLY SPLIT-AND-REPHRASING TECHNIQUES

	Approach	Language	Split Technique	Rephrase Technique	Transformer Architecture
[8]	Context-Preserving Text Simplification	English	Defining a 35 hand-crafted transformation rules	Semantic hierarchy of minimal propositions	✗
[9]	Fact-Aware Sentence Split and Rephrase with Permutation Invariant Training	English	Training a CNN model for predicting sentence split	Seq2seq Model	✗
[7]	Controllable Text Simplification with Explicit Paraphrasing	English	DisSim: structural simplification tool consisting of 35 hand-crafted grammar rules	Transformer seq2seq Model	✓
[10]	A Memory-Based Sentence Split and Rephrase Model	English	Encoder and Decoder transformer model connected by a memory gate layer		✓
[20]	Hierarchical Generation for Sentence Simplification	English	A semantic separator layer at decoder model	LSTM Seq2seq based Model	✗
	our TSimAr	Arabic	PDTS [5] (built upon mBERT model)	FNet Text-to-Text based model	✓

two main steps (i.e. the segmentation and rephrase steps) in a little more detail.

### A. Text Segmentation

We base our proposed TSimAr on top of PDTS [5] (i.e. an Arabic text splitting tool that employs a pre-trained multilingual BERT [55] model for detecting missing punctuations) for segmenting input texts into a set of potentially independent-clauses. More in detail, PDTS queries mBERT<sup>5</sup>  $|X|$  times to predict proper punctuations between words on which they can be used as text split delimiters:

$$p(pu_i^m | t_i^m; pun, \theta) = PDTS \circ mBERT(X'_i), \quad (1)$$

$$X'_i \leftarrow insertMaskToken(X, i), \forall i \in X \quad (2)$$

where  $pu_i^m$  represents the valid mBERT's output for a given white-space at index  $i$ ;  $t_i^m$  is the actual mBERT's output;  $pun$  and  $\theta$  are model parameters set by the user to filter out  $t_i^m$ ; and  $X'_i$  is the input  $X$  with the inserted  $[MASK]$  token at index  $i$ . PDTS then validates the predicted set of punctuations  $pu^m \forall i \in \{1, 2, \dots, |X|\}$  using four generic linguistic rules in a greedy-like strategy. In this paper, we have set  $pun$  with only the main splitting punctuations, including full-stop, comma, semicolon, and colon.

### B. Rephrase Generation

Motivated by Transformers-based encoder-decoder architecture [56] that has achieved outstanding improvements in complicated NLP tasks, we consider one of its optimized sequence-to-sequence models. In particular, we utilize (FNet) an efficient version that substitutes complex self-attention layers with linear Fourier Transforms-based layers, introduced originally in [21]. Here, FNet is efficiently lighter and much faster than the standard Transformer model with complicated attention layers. Besides, it closely matches the performance of the standard Transformer model. Considering the right part of Figure 2, it describes the core part of standard Transformer architecture that is modified by adding fast Fourier Transforms layers. Broadly speaking, *Transformer* blocks are stacked with a size of  $N_x$ , where each block consists of a two residual

gateless layers that adds additional weight matrix with *Skip Connection*, Eq. (3):

$$y := \epsilon f(x) + x \quad (3)$$

where  $\epsilon$  is the regularization parameter. In the standard Transformer architecture, the multi-head attention (i.e. concatenate a number of self-attention layers) allows to learn the structural and morphological correlation between different input-tokens impressively, using Eq. (4):

$$attention(Q_i, K_i, V_i) = \sigma \left( \frac{Q_i K_i^T}{\sqrt{d_k}} \right) V_i, \quad \text{for } i = 1, \dots, h \text{ (head)}. \quad (4)$$

where **query**, **key**, and **value** vectors are computed as  $Q_i = W_q^i x$ ,  $K_i = W_k^i x$ ,  $V_i = W_v^i x$ ; and  $\sigma$  is the softmax activation function. Nevertheless, it is memory intensive and has an exponential (quadratic) time complexity concerning the size of the input sequence [56]. Thus, to avoid such scalability issues, we employ fast Fourier Transform layers as an alternative to attention sublayers, expressed in Eq. (5):

$$W_{ij} = \left( \frac{e^{-i \frac{2\pi}{N} i \times j}}{\sqrt{N}} \right) \quad (5)$$

where  $i, j = 0, \dots, N - 1$ . Recent experiments [21] demonstrated that the Fourier Transform-based model (so-called FNet) can significantly reduce the training time and space complexity while providing an excellent performance that is exceptionally comparable to the performance offered by the standard Transformer-based encoder-decoder model. The architecture of our implemented FNet sequence-to-sequence model for rephrasing Arabic texts is presented explicitly in Figure 3.

## IV. EXPERIMENTS AND RESULTS

We conduct experiments to assess the performance of TSimAr and analogize its rephrasing part with the existing Arabic pre-trained text-to-text generation models. We introduce ATSC (a new Arabic corpus for text simplification) and use it in our evaluation protocols. As a quality benchmark of generated simplifications, we report various automatic text

<sup>5</sup><https://github.com/google-research/bert/blob/master/multilingual.md>

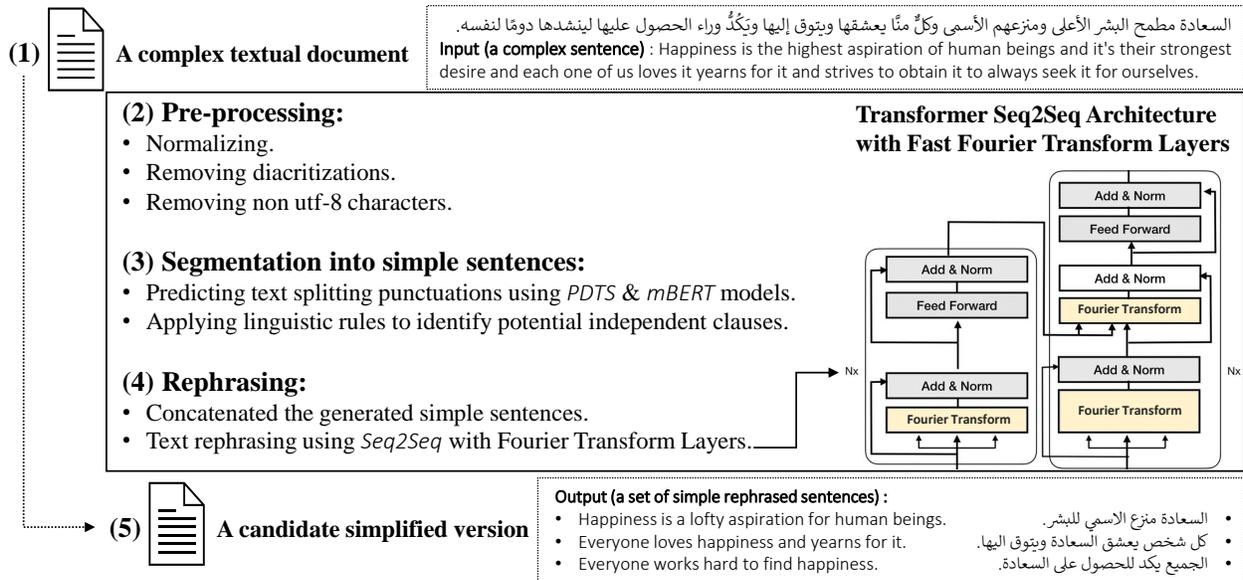


Fig. 2. The overview flow of the proposed TSimAr with an illustrative input-output example

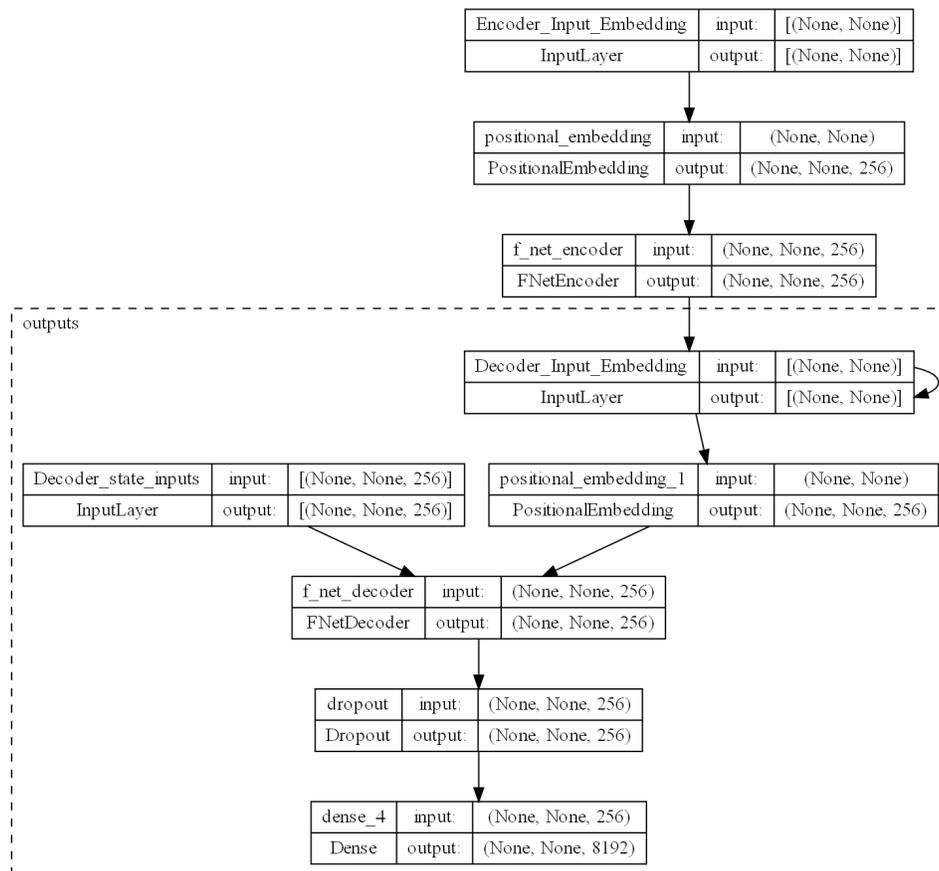


Fig. 3. The architecture of our implemented FNet sequence-to-sequence model for Arabic texts rephrasing. The total trainable parameters are 11,055,616

matching metrics, including SARI (the primary metric for text simplification and rephrasing), besides presenting the findings from the conducted manual (human-based) assessment.

#### A. Corpus and Experimental Setup

a) Arabic text simplification corpus (ATSC):: To the extent of our knowledge, there is no specific Arabic corpus for

text simplification. Thus, we create a small benchmark corpus containing 500 pairs of a small-to-large complex text (source) and a gold-standard simplified (i.e. splitted and rephrased) text. The gold-standard reference simplifications are written and carefully reviewed by human experts. In a little more detail, our corpus ATSC has been constructed from selective Arabic articles that contain appropriate text to simplify. We collected these articles from different public sources (i.e., Wikipedia, newspapers, and news agencies), which cover various domains, including history, geography, health, education, and technology. For constructing the simplified versions from the collected articles (i.e., form the gold-standard human-based references), we have applied two simplification methods: *syntactic simplification* (i.e. just an extractive text summarization method that drops/selects the key sub-sentences without generating new words) as well as *linguistic simplification* (i.e. almost similar to the abstractive summarization method that attempts to replace complicated words with conceivably simpler synonym words), depending on their contexts and overall meaning. Table II shows some general statistical descriptions of our ATSC.

TABLE II. STATISTICAL DESCRIPTIONS OF OUR EVALUATION CORPUS (ATSC)

	Original complex texts	Simplified Rewrites
# No. Documents	500	500
# No. Sentences	546	1243
# No. Words (distinct)	17069 (6046)	17564 (5479)
# No. Vocabularies (for both complex and simplified texts)	6737	

*b) Baselines::* We compare the text rephrasing part in our TSimAr (i.e. FNet model) against the state-of-the-art pre-trained Arabic monolingual (Arabic-T5-small [16], Arabic-T5 [15], UBC-AraT5 [17]) and general multilingual (MT5-base [19], mBART-large-50 [18]) models for text generation tasks. These text-to-text generation models are architecturally extended from T5 encoder-decoder transformer blocks[57], except mBART that is a multilingual Sequence-to-Sequence model used generally for translation tasks:

*c) Automatic metrics::* Given a source text  $st$  and a gold-reference  $gr$  (i.e. a typical simplification version written by human experts), we evaluate the efficiency of the produced simplification  $y$  (i.e.  $Y \leftarrow TSimAr(st)$ ) using a variety of automatic metrics as follows:

- SARI [58] (*System output against References and against the Input sentence*) is a standard evaluation metric for text simplification, which compares the generated candidate simplifications  $y$  against both (1) the source input  $st$  and (2) the gold-reference  $gr$ . It uses precision and F1 scores of  $n$ -grams ( $n \in 1, 2, 3, 4$ ) to measure the goodness of added, deleted, and preserved tokens by the simplifier model (i.e. TSimAr).
- BLEU [59] (Bilingual Evaluation Understudy) is a popular evaluation metric for text quality, commonly used in machine-translated tasks. It compares  $y$  against  $gr$  only and approximates recall and precision metrics using the best match ( $n$ -gram) length and modified  $n$ -gram precision, respectively.
- METEOR [60] (*Metric for Evaluation for Translation with Explicit Ordering*) is similar to BLEU but replaces the best match ( $n$ -gram) length and modified  $n$ -gram

precision with a weighted F-score metric that depends on unigram mapping.

- TER [61] (Translation Edit Rate) which estimates the number of edits required (e.g., adding, deleting, or shifting a word token) to improve  $y$  as matched with  $gr$ .
- ROUGE [36] (Recall-Oriented Understudy for Gisting Evaluation) gives different ROUGE- $n$  metrics, where  $n$  represents the number of overlapping  $n$ -grams between  $y$  and  $gr$ . It uses the standard statistical metrics (precision, recall, and F-measure) for its measurements. In our experiments, we consider ROUGE-1 (unigram overlapping), ROUGE-2 (bi-grams overlapping), and ROUGE-L (the longest identical subsequence overlapping between  $y$  and  $gr$ ).

Concerning SARI, BLEU, METEOR, and ROUGE, higher scores indicate better quality correlated to rational human judgments. In contrast, a lower TER metric (i.e. lower editing scores) indicates better performance.

*d) Implementation details:* To train and configure the text rephrasing part in our TSimAr (presented in Figure 3), we applied a 50—20—30 random split on our ATSC corpus to create train, dev, and test sets, respectively. Besides, we used the Adam optimization algorithm for training with a learning rate of 0.001. The training loop lasts 5k epochs with a batch size of 64 and a maximum sequence length of 256. Moreover, the text-to-text generation models, considered in this paper as baselines, are publicly available at the Hugging Face<sup>6</sup>, under the model (card) names: ‘google/mt5-base’, ‘facebook/mbart-large-50’, ‘flax-community/arabic-t5-small’, ‘UBC-NLP/AraT5-base-title-generation’, and ‘malmarjeh/t5-arabic-text-summarization’. We have constructed these models using PyTorch<sup>7</sup> framework besides utilizing some NLP toolkits for text preprocessing, including NLTK<sup>8</sup> and CAMEL<sup>9</sup>. All experiments have been conducted using a gaming PC equipped with Intel i9-CPU, 64G-RAM, and a single NVIDIA GeForce RTX3070 GPU.

## B. Performance Evaluation

In Table III, we show the performance of our TSimAr with FNet against different text rephrasing models (i.e. depending on text-to-text generation models) using the validation portion from ATSC. Performance results are also visualized in Figure 4. Besides that, we break down the performance details and simplification quality for one input instance in Table IV. As can be observed, TSimAr evidently outperforms all the existing state of the art text-to-text generation models for the Arabic language. It achieves the best score on all standard metrics (particularly SARI) and gives the second to the best score on TER and ROUGE-1. The last column on the right of Table III shows the execution time in second, visualized in Figure 5. Here, our TSimAr gives foreseen poor-to-ordinary time performance as its FNet architecture is quite heavy (consisting of more than 11M trainable parameters).

In addition, giving insight into the text simplifications produced by the competitors’ models, presented in Table IV

<sup>6</sup><https://huggingface.co/>

<sup>7</sup><https://pytorch.org>

<sup>8</sup><https://www.nltk.org/>

<sup>9</sup><https://camel-tools.readthedocs.io/>

TABLE III. AUTOMATIC EVALUATION RESULTS. THE BEST PERFORMANCE FOUND IS INDICATED BY THE ASTERISK\*

	SARI	BLEU	TER	METEOR	ROUGE			ET. (Sec)
					R-1	R-2	R-L	
Arabic-T5-small [16]	0.23	0.02	0.89	0.11	0.23	0.11	0.22	128 *
Arabic-T5 [15]	0.21	0.01	0.94	0.05	0.14	0.04	0.14	145
MT5-base [19]	0.18	0.00	0.98 *	0.02	0.05	0.01	0.04	213
mBART [18]	0.53	0.40	0.52	0.61	0.72 *	0.53	0.70	3685
UBC-AraT5 [17]	0.21	0.00	0.95	0.05	0.12	0.04	0.11	178
Our TSimAr	0.73 *	0.65 *	0.34	0.68 *	0.71	0.67 *	0.71 *	724

TABLE IV. EXAMPLES OF GENERATED TEXT SIMPLIFICATIONS BY RELATED BASELINE MODELS, INCLUDING OURS. THE INPUT EXAMPLE IS TAKEN RANDOMLY FROM THE VALIDATION SET, AND THE SIMPLIFICATIONS' QUALITY IS ILLUSTRATED USING OUR CONSIDERED AUTOMATIC METRICS

**Input: complex sentences**

ويستدعي التعامل مع الاحتياجات الصحية المتزايدة التعقيد أتباع نهج متعدد القطاعات تُدمج فيه السياسات الهادفة لتعزيز الصحة والوقاية، مع وضع حلول تستجيب للمجتمعات المحلية، وتقديم خدمات صحية تتمحور حول الناس. وتشمل الرعاية الصحية الأولية أيضاً العناصر الأساسية اللازمة لتحسين الأمن الصحي ودرء التهديدات الصحية مثل الأوبئة ومقاومة مضادات الميكروبات، من خلال تدابير مثل المشاركة المجتمعية والتعليم، والوصفات الطبية الرشيدة، ومجموعة أساسية من وظائف الصحة العمومية الأساسية، بما في ذلك الترصد.

Dealing with increasingly complex health needs calls for a multisectoral approach in which health promotion and prevention policies are combined, with community-based solutions and people-centred health services. Primary health care also includes the essential elements needed to improve health security and stave off health threats such as epidemics and antimicrobial resistance, through measures such as community participation and education, rational prescription, and a core set of essential public health functions, including surveillance.

Model	Generated text
Arabic-T5-small	Including vital goals, and making provisions بما في ذلك الاهداف الحيويه ، واجراء احكام
Arabic-T5	"الصحة الصحية" تواجه الاحتياجات الصحية المتزايدة المعقدة
MT5-base	"Healthcare" is facing increasingly complex health needs
MT5-base	It can also كما يمكن <extra_id_0>
mBART-large-50	ويستدعي التعامل مع الاحتياجات الصحية المتزايدة التعقيد أتباع نهج متعدد القطاعات تدمج فيه السياسات الهادفة لتعزيز الصحة والوقاية ، مع وضع حلول تستجيب للمجتمعات المحلية ، وتقديم خدمات صحية تتمحور حول الناس . وتشمل الرعاية الصحية الأولية ايضا العناصر الاساسيه اللازمه لتحسين الامن الصحي ودرء التهديدات الصحيه مثل الاوبئه ومقاومه مضادات الميكروبات ، من خلال تدابير مثل المشاركة المجتمعيه والتعليم ، والوصفات الطبيه الرشيدته ، ومجموعه اساسيه من وظائف الصحة العموميه الاساسيه ، بما في ذلك الترصد . . .
UBC-AraT5	وزارة الصحة: الدعم الصحي هو الحل الأمثل لمواجهة الاحتياجات الصحية
Our model	Ministry of Health: Health support is the best solution to meet health needs للتعامل مع الاحتياجات الصحية المعقدة يجب ان نتبع نهج متعدد القطاعات لتعزيز الصحة . يجب تقديم خدمات صحيه تتمحور حول الناس . العناصر الاساسيه للرعايه الصحيه الاوليه هي المشاركة المجتمعيه ، الوصفات الطبيه الرشيدته والوظائف الاساسيه للصحة العامه . To deal with complex health needs we must take a multi-sectoral approach to health promotion. People-centred health services must be provided. The essential elements of primary health care are community participation, rational prescriptions, and basic functions of public health.

TABLE V. AUTOMATIC EVALUATION RESULTS FOR THE TEXT EXAMPLE DESCRIBED IN TABLE IV

	SARI	BLEU	TER	METEOR	ROUGE		
					R-1	R-2	R-L
Arabic-T5-small [16]	0.27	0.00	0.98	0.01	0.04	0.00	0.04
Arabic-T5 [15]	0.27	0.00	1.00	0.01	0.04	0.00	0.04
MT5-base [19]	0.27	0.00	0.98	0.01	0.05	0.00	0.05
mBART [18]	0.44	0.15	1.10	0.57	0.51	0.28	0.50
UBC-AraT5 [17]	0.27	0.00	1.00	0.01	0.04	0.00	0.04
Our TSimAr	0.90 *	0.79 *	0.19	0.81 *	0.92 *	0.87 *	0.92 *

and Table V, one can observe that mid-to-high automatic metrics results may not necessarily reflect valid candidate

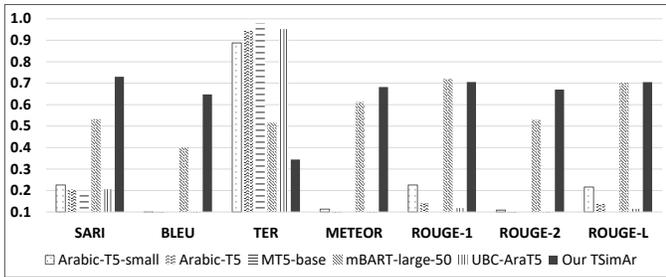


Fig. 4. Performance of TSimAr with FNet against competitors' text rephrasing models (i.e. depending on text-to-text generation models)

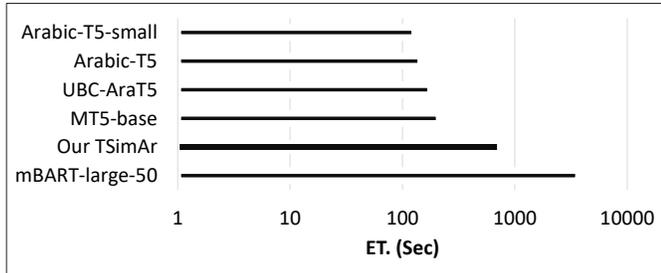


Fig. 5. Performance (w.r.t execution time in second) of TSimAr with FNet against competitors' text rephrasing models

simplification. To clarify more, we observe that mBART [18] often produces outputs almost identical to the input without simplifying or rephrasing, and in turn, it inaccurately archives more than 0.4 SARI score. While, UBC-AraT5 [17] simplifies input text in much better quality, it achieves around 0.27 SARI. Accordingly, it was essential to solidify the evaluation of our proposed TSimAr using manual insight by eliciting human judgments.

### C. Manual Evaluation

To get an additional in-depth evaluation of our TSimAr, we conducted a qualitative analysis by eliciting a humanistic viewpoint on 36 sampled text documents selected randomly from the ATSC validation set. We invited two expert consultants in Arabic linguistics (not authors of this paper) to evaluate these documents (each expert is given 18 documents) on the following three standards using a five-star-point Likert scale (1-5):

- Adequacy (preservation of the source meaning),
- Contextual soundness (quality of rephrased and simplified texts), and
- Grammaticality (to what extent the generated text is free from grammatical errors).

Experts are asked to compare the generated simplifications TSimAr by (i.e. depending on Arabic-T5-small, Arabic-T5, MT5-base, mBART, UBC-AraT5, and our FNet) against the gold-standard references (i.e. text simplification versions written by human experts). With a glancing over into Table VI, the results of our manual evaluation look almost compatible with the automatic evaluation results (shown in Table III) for only the first and the third standards (i.e. *Adequacy* or *Grammaticality*). Nevertheless, *Contextual soundness* standard

TABLE VI. HUMAN EVALUATION RESULTS FOR THE THREE CRITERIA: ADEQUACY, CONTEXTUAL SOUNDNESS, AND GRAMMATICALITY. BASE PRE-TRAINED TEXT-TO-TEXT GENERATION MODELS WITH \* ARE SIGNIFICANTLY DIFFERENT FROM TSIMAR'S BASE MODEL †, DEPENDING ON A TWO-TAILED INDEPENDENT T-TEST, WHERE  $p < .05$ . FOR ARABIC-T5, THE DIFFERENCE IS SIGNIFICANT AT ( $p < .1$ )

Text generation models	A	C	G	Avg.	p-value	t-value
Arabic-T5-small [16]	1.5	2	3	2.17	.010585 *	4.52904
Arabic-T5 [15]	3	3.5	4.5	3.67	.080472 *	1.71791
MT5-base [19]	0	0	1	0.33	.000694 *	9.47046
mBART [18]	5	0	4	3.00	.366645	1.01705
UBC-AraT5 [17]	3.5	4.75	5	4.42	.778051	0.30151
Our TSimAr depending on FNet	4	4.75	5	4.58	†	

reveals the quality differences more precisely, which also confirms that our TSimAr with FNet (indicated by †) can produce a highly competitive performance (see, 4.58 the best average ratings obtained by TSimAr).

Furthermore, the manual experimental results indicated that UBC-AraT5 is a feasible pre-trained text rephrasing model to adopt (i.e. an alternative model to FNet) as it achieves the second highest average score of about 4.42. This indication can also be statistically noticed by its insignificant p-value (i.e. the obtained .78 with UBC-AraT5). In contrast, however, the worst performances observed were with MT5-base and mBART, which unexpectedly gave zero simplification quality. Here, the performance of mBART contradicts the insignificant p-value (i.e. .367) as this heavy model offers an illusive high score in *Adequacy* and *Grammaticality* standards, which is a consequence of generating texts exactly similar to the input texts with no simplification.

### D. Discussion and Potential Threats to Validity

In this section, we discuss the potential threats to the empirical validity of the proposed TSimAr. The main threats may include the creation of our corpus (ATSC) for evaluation as well as the benchmarking against the state-of-the-art text rephrasing models. As mentioned earlier in this section, there is no specific Arabic corpus for text simplification available to date. Therefore, we had to make an effort to (1) collect professionally written corpus from online sources (mainly from newspaper articles) and (2) simplify them precisely (i.e. split with rephrasing) by linguistic experts in Arabic, elaborated in ATSC. One may argue that ATSC is relatively small (containing only 500 pairs of texts), and more importantly, it may be insufficient to train a heavy model like FNet. These thoughts are valid to a reasonable extent. However, recent studies demonstrated that training a language understanding model on a larger corpora/dataset might not necessarily imply improving its performance [62], [63]. Besides, our intention here is not to use ATSC to train a language understating model but rather to use it as a benchmarking corpus for testing the generalization of a pre-trained test-to-text generation model. We make our ATSC available for researchers to exploit in this regard.

Concerning the chosen text rephrasing baseline models, we attempted to counter this concern by using all publicly available Arabic monolingual sequence-to-sequence models (we have encountered only Arabic-T5, Arabic-T5-small, and UBC-AraT5) as well as using the state-of-the-art multilingual models

(i.e. MT5 and mBART). For a fair comparison between these pre-trained models, we confirmed that their large vocabulary contains all the distinct 6737 words extracted from ATSC.

## V. CONCLUSION

Approaching towards breaking down a given complex Arabic text into a simple and meaning-preserving version, we have presented a text split-and-rephrase solution (so-called TSimAr), which depends principally on a sequence-to-sequence Transformer-based architecture. For the splitting, we have integrated TSimAr with a punctuation detector for text segmentation (PDTS) built on top of a pre-trained multilingual masked-language model (mBERT). This PDTS attempts to generate the shortest set of simple independent-clause sentences from a given lengthy complex text. While in the rephrasing phase, we have proposed an attention-free Transformer model, depending on a fast Fourier-Transform (FNet-based), which rephrases the concatenated simple sentences into a more readable version.

In addition, we have created a new corpus (ATSC) to train and evaluate the rephrasing part in our TSimAr. Automated and manual analyses demonstrated that with the support of PDTS, our TSimAr evidently outperforms all the existing state-of-the-art text-to-text generation models for the Arabic language as it achieved the best score on SARI, BLEU, and METEOR metrics. Nevertheless, a trivial limitation noted in TSimAr lies in the execution time compared with competitors' lighter models, such as Arabic-T5-small. Hence, for the generality, we imagine a remarkable extension of this ongoing work in two directions:

- (1) evaluating TSimAr on a comprehensively benchmarking dataset that we plan to create, and
- (2) optimizing our FNet architecture for enhancing its execution performance.

For the latter direction, we will investigate the feasibility of applying a knowledge distillation technique to compress our FNet into a smaller version to help us reduce its space complexity while achieving higher inference speed and accuracy.

## ACKNOWLEDGMENT

The authors extend their appreciation to the Deputyship for Research & Innovation at the Ministry of Education in the Kingdom of Saudi Arabia for funding this research work through project number (19/20).

## REFERENCES

[1] O. Alonzo, "The use of automatic text simplification to provide reading assistance to deaf and hard-of-hearing individuals in computing fields," *SIGACCESS Access. Comput.*, vol. 3, no. 132, mar 2022. [Online]. Available: <https://doi-org.sdl.idm.oclc.org/10.1145/3523265.3523268>

[2] T. Hao, X. Li, Y. He, F. L. Wang, and Y. Qu, "Recent progress in leveraging deep learning methods for question answering," *Neural Computing and Applications*, vol. 34, pp. 1–19, 2022.

[3] A. Alshantqi, A. Namoun, A. Alsughayyir, A. M. Mashraqi, A. R. Gilal, and S. S. Albouq, "Leveraging distilbert for summarizing arabic text: An extractive dual-stage approach," *IEEE Access*, vol. 9, pp. 135 594–135 607, 2021.

[4] F. Alva-Manchego, C. Scarton, and L. Specia, "Data-driven sentence simplification: Survey and benchmark," *Computational Linguistics*, vol. 46, no. 1, pp. 135–187, 2020.

[5] A. M. Alshantqi, S. Albouq, A. B. Alkhodre, A. Namoun, and E. Nabil, "Employing a multilingual transformer model for segmenting unpunctuated arabic text," *Applied Sciences*, vol. 12, no. 20, 2022. [Online]. Available: <https://www.mdpi.com/2076-3417/12/20/10559>

[6] C. Niklaus, M. Cetto, A. Freitas, and S. Handschuh, "Transforming complex sentences into a semantic hierarchy," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, July 2019, 57th Annual Meeting of the Association for Computational Linguistics ; Conference date: 28-07-2019 Through 02-08-2019.

[7] M. Maddela, F. Alva-Manchego, and W. Xu, "Controllable text simplification with explicit paraphrasing," *arXiv preprint arXiv:2010.11004*, 2020.

[8] C. Niklaus, M. Cetto, A. Freitas, and S. Handschuh, "Context-preserving text simplification," *arXiv preprint arXiv:2105.11178*, 2021.

[9] Y. Guo, T. Ge, and F. Wei, "Fact-aware sentence split and rephrase with permutation invariant training," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, 2020, pp. 7855–7862.

[10] X. Fan, Y. Liu, G. Liu, and B. Su, "A memory-based sentence split and rephrase model with multi-task training," in *International Conference on Neural Information Processing*. Springer, 2020, pp. 643–654.

[11] L. Martin, A. Fan, É. de la Clergerie, A. Bordes, and B. Sagot, "Muss: multilingual unsupervised sentence simplification by mining paraphrases," *arXiv preprint arXiv:2005.00352*, 2020.

[12] D. Gamal, M. Alfonse, S. M. Jiménez-Zafra, and M. Aref, "Survey of arabic machine translation, methodologies, progress, and challenges," in *2022 2nd International Mobile, Intelligent, and Ubiquitous Computing Conference (MIUCC)*. IEEE, 2022, pp. 378–383.

[13] M. Zhou, N. Duan, S. Liu, and H.-Y. Shum, "Progress in neural nlp: Modeling, learning, and reasoning," *Engineering*, vol. 6, no. 3, pp. 275–290, 2020. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S2095809919304928>

[14] P. Sikka and V. Mago, "A Survey on Text Simplification," May 2022, arXiv:2008.08612 [cs]. [Online]. Available: <http://arxiv.org/abs/2008.08612>

[15] F. Community, "Arabic-t5 (small)," 2022. [Online]. Available: <https://huggingface.co/malmarjeh/t5-arabic-text-summarization>

[16] M. B. Almarjeh, "An arabic abstractive text summarization model," 2022. [Online]. Available: <https://huggingface.co/flax-community/arabic-t5-small>

[17] E. M. B. Nagoudi, A. Elmadany, and M. Abdul-Mageed, "AraT5: Text-to-text transformers for Arabic language generation," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland: Association for Computational Linguistics, may 2022, pp. 628–647. [Online]. Available: <https://aclanthology.org/2022.acl-long.47>

[18] Y. Tang, C. Tran, X. Li, P.-J. Chen, N. Goyal, V. Chaudhary, J. Gu, and A. Fan, "Multilingual translation with extensible multilingual pretraining and finetuning," *arXiv eprint 2008.00401*, 2020.

[19] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, and C. Raffel, "mt5: A massively multilingual pre-trained text-to-text transformer," *CoRR*, vol. abs/2010.11934, 2020. [Online]. Available: <https://arxiv.org/abs/2010.11934>

[20] M. Wang, H. Ozaki, Y. Koreeda, and K. Yanai, "Split first and then rephrase: Hierarchical generation for sentence simplification," in *International Conference of the Pacific Association for Computational Linguistics*. Springer, 2019, pp. 15–27.

[21] I. Eckstein, J. P. Lee-Thorp, J. Ainslie, and S. Ontanon, Eds., *FNet: Mixing Tokens with Fourier Transforms*, 2022. [Online]. Available: <https://aclanthology.org/2022.naacl-main.319.pdf>

[22] E. Pavlick and C. Callison-Burch, "Simple PPDB: A Paraphrase Database for Simplification," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 143–148. [Online]. Available: <https://aclanthology.org/P16-2024>

[23] M. Maddela and W. Xu, "A Word-Complexity Lexicon and A Neural Readability Ranking Model for Lexical Simplification," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for

- Computational Linguistics, Oct. 2018, pp. 3749–3760. [Online]. Available: <https://aclanthology.org/D18-1410>
- [24] M. Yatskar, B. Pang, C. Danescu-Niculescu-Mizil, and L. Lee, “For the sake of simplicity: Unsupervised extraction of lexical simplifications from Wikipedia,” in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Los Angeles, California: Association for Computational Linguistics, Jun. 2010, pp. 365–368. [Online]. Available: <https://aclanthology.org/N10-1056>
- [25] C. Horn, C. Manduca, and D. Kauchak, “Learning a Lexical Simplifier Using Wikipedia,” in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Baltimore, Maryland: Association for Computational Linguistics, Jun. 2014, pp. 458–463. [Online]. Available: <https://aclanthology.org/P14-2075>
- [26] O. Biran, S. Brody, and N. Elhadad, “Putting it Simply: a Context-Aware Approach to Lexical Simplification,” in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Portland, Oregon, USA: Association for Computational Linguistics, Jun. 2011, pp. 496–501. [Online]. Available: <https://aclanthology.org/P11-2087>
- [27] G. Glavaš and S. Štajner, “Simplifying Lexical Simplification: Do We Need Simplified Corpora?” in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Beijing, China: Association for Computational Linguistics, Jul. 2015, pp. 63–68. [Online]. Available: <https://aclanthology.org/P15-2011>
- [28] G. Paetzold and L. Specia, “Lexical Simplification with Neural Ranking,” in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. Valencia, Spain: Association for Computational Linguistics, Apr. 2017, pp. 34–40. [Online]. Available: <https://aclanthology.org/E17-2006>
- [29] S. Gooding and E. Kochmar, “Recursive Context-Aware Lexical Simplification,” in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 4853–4863. [Online]. Available: <https://aclanthology.org/D19-1491>
- [30] J. Qiang, Y. Li, Y. Zhu, Y. Yuan, and X. Wu, “Lexical Simplification with Pretrained Encoders,” Oct. 2020, arXiv:1907.06226 [cs]. [Online]. Available: <http://arxiv.org/abs/1907.06226>
- [31] W. Zhou, T. Ge, K. Xu, F. Wei, and M. Zhou, “BERT-based Lexical Substitution,” in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 3368–3373. [Online]. Available: <https://aclanthology.org/P19-1328>
- [32] S. Aluísio and C. Gasperin, “Fostering Digital Inclusion and Accessibility: The PorSimples project for Simplification of Portuguese Texts,” in *Proceedings of the NAACL HLT 2010 Young Investigators Workshop on Computational Approaches to Languages of the Americas*. Los Angeles, California: Association for Computational Linguistics, Jun. 2010, pp. 46–53. [Online]. Available: <https://aclanthology.org/W10-1607>
- [33] S. Jonnalagadda and G. Gonzalez, “BioSimplify: an open source sentence simplification engine to improve recall in automatic biomedical information extraction,” *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, vol. 2010, pp. 351–5, Nov. 2010.
- [34] N. Bach, Q. Gao, S. Vogel, and A. Waibel, “TriS: A Statistical Sentence Simplifier with Log-linear Models and Margin-based Discriminative Training,” in *Proceedings of 5th International Joint Conference on Natural Language Processing*. Chiang Mai, Thailand: Asian Federation of Natural Language Processing, Nov. 2011, pp. 474–482. [Online]. Available: <https://aclanthology.org/I11-1053>
- [35] M. Heilman and N. A. Smith, “Extracting Simplified Statements for Factual Question Generation.”
- [36] C.-Y. Lin, “ROUGE: A Package for Automatic Evaluation of Summaries,” in *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics, Jul. 2004, pp. 74–81. [Online]. Available: <https://aclanthology.org/W04-1013>
- [37] G. H. Paetzold and L. Specia, “Text Simplification as Tree Transduction,” in *Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology*, 2013. [Online]. Available: <https://aclanthology.org/W13-4813>
- [38] M. Shardlow, “A Survey of Automated Text Simplification,” *International Journal of Advanced Computer Science and Applications*, vol. 4, no. 1, Apr. 2014, number: 1 Publisher: The Science and Information (SAI) Organization Limited. [Online]. Available: <https://thesai.org/Publications/ViewPaper?Volume=4&Issue=1&Code=SpecialIssue&SerialNo=9>
- [39] W. Coster and D. Kauchak, “Learning to Simplify Sentences Using Wikipedia,” in *Proceedings of the Workshop on Monolingual Text-To-Text Generation*. Portland, Oregon: Association for Computational Linguistics, Jun. 2011, pp. 1–9. [Online]. Available: <https://aclanthology.org/W11-1601>
- [40] H.-B. Chen, H.-H. Huang, H.-H. Chen, and C.-T. Tan, “A Simplification-Translation-Restoration Framework for Cross-Domain SMT Applications,” in *Proceedings of COLING 2012*. Mumbai, India: The COLING 2012 Organizing Committee, Dec. 2012, pp. 545–560. [Online]. Available: <https://aclanthology.org/C12-1034>
- [41] D. Klaper, S. Ebling, and M. Volk, “Building a German/Simple German Parallel Corpus for Automatic Text Simplification,” in *Proceedings of the Second Workshop on Predicting and Improving Text Readability for Target Reader Populations*. Sofia, Bulgaria: Association for Computational Linguistics, Aug. 2013, pp. 11–19. [Online]. Available: <https://aclanthology.org/W13-2902>
- [42] S. Stymne, J. Tiedemann, C. Hardmeier, and J. Nivre, “Statistical Machine Translation with Readability Constraints,” in *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013)*. Oslo, Norway: Linköping University Electronic Press, Sweden, May 2013, pp. 375–386. [Online]. Available: <https://aclanthology.org/W13-5634>
- [43] K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, “Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation,” in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 1724–1734. [Online]. Available: <https://aclanthology.org/D14-1179>
- [44] T. Wang, P. Chen, J. Rochford, and J. Qiang, “Text simplification using Neural Machine Translation,” in *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, ser. AAAI’16. Phoenix, Arizona: AAAI Press, Feb. 2016, pp. 4270–4271.
- [45] T. Wang, P. Chen, K. Amaral, and J. Qiang, “An Experimental Study of LSTM Encoder-Decoder Model for Text Simplification,” Sep. 2016, arXiv:1609.03663 [cs]. [Online]. Available: <http://arxiv.org/abs/1609.03663>
- [46] S. Wubben, A. van den Bosch, and E. Krahmer, “Sentence Simplification by Monolingual Machine Translation,” in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Jeju Island, Korea: Association for Computational Linguistics, Jul. 2012, pp. 1015–1024. [Online]. Available: <https://aclanthology.org/P12-1107>
- [47] Z. Zhu, D. Bernhard, and I. Gurevych, “A Monolingual Tree-based Translation Model for Sentence Simplification,” in *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*. Beijing, China: Coling 2010 Organizing Committee, Aug. 2010, pp. 1353–1361. [Online]. Available: <https://aclanthology.org/C10-1152>
- [48] K. Woodsend and M. Lapata, “Learning to Simplify Sentences with Quasi-Synchronous Grammar and Integer Programming,” in *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*. Edinburgh, Scotland, UK.: Association for Computational Linguistics, Jul. 2011, pp. 409–420. [Online]. Available: <https://aclanthology.org/D11-1038>
- [49] A. See, P. J. Liu, and C. D. Manning, “Get To The Point: Summarization with Pointer-Generator Networks,” Apr. 2017, arXiv:1704.04368 [cs]. [Online]. Available: <http://arxiv.org/abs/1704.04368>
- [50] X. Zhang and M. Lapata, “Sentence Simplification with Deep Reinforcement Learning,” in *Proceedings of the 2017 Conference on*

- Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, Sep. 2017, pp. 584–594. [Online]. Available: <https://aclanthology.org/D17-1062>
- [51] S. Štajner and H. Saggion, “Data-Driven Text Simplification,” in *Proceedings of the 27th International Conference on Computational Linguistics: Tutorial Abstracts*. Santa Fe, New Mexico, USA: Association for Computational Linguistics, Aug. 2018, pp. 19–23. [Online]. Available: <https://aclanthology.org/C18-3005>
- [52] E. Sulem, O. Abend, and A. Rappoport, “Semantic Structural Evaluation for Text Simplification,” Oct. 2018, arXiv:1810.05022 [cs]. [Online]. Available: <http://arxiv.org/abs/1810.05022>
- [53] J. Qiang and X. Wu, “Unsupervised Statistical Text Simplification,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 33, no. 4, pp. 1802–1806, Apr. 2021, conference Name: IEEE Transactions on Knowledge and Data Engineering.
- [54] S. Surya, A. Mishra, A. Laha, P. Jain, and K. Sankaranarayanan, “Unsupervised Neural Text Simplification,” in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 2058–2068. [Online]. Available: <https://aclanthology.org/P19-1198>
- [55] 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.
- [56] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” in *Advances in neural information processing systems*, 2017, pp. 5998–6008.
- [57] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu *et al.*, “Exploring the limits of transfer learning with a unified text-to-text transformer.” *J. Mach. Learn. Res.*, vol. 21, no. 140, pp. 1–67, 2020.
- [58] W. Xu, C. Napoles, E. Pavlick, Q. Chen, and C. Callison-Burch, “Optimizing statistical machine translation for text simplification,” *Transactions of the Association for Computational Linguistics*, vol. 4, pp. 401–415, 2016.
- [59] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, “Bleu: a method for automatic evaluation of machine translation,” in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 2002, pp. 311–318.
- [60] S. Banerjee and A. Lavie, “METEOR: An automatic metric for MT evaluation with improved correlation with human judgments,” in *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*. Ann Arbor, Michigan: Association for Computational Linguistics, Jun 2005, pp. 65–72. [Online]. Available: <https://www.aclweb.org/anthology/W05-0909>
- [61] M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul, “A study of translation edit rate with targeted human annotation,” in *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*. Cambridge, Massachusetts, USA: Association for Machine Translation in the Americas, Aug 2006, pp. 223–231. [Online]. Available: <https://aclanthology.org/2006.amta-papers.25>
- [62] G. Inoue, B. Alhafni, N. Baimukan, H. Bouamor, and N. Habash, “The interplay of variant, size, and task type in Arabic pre-trained language models,” in *Proceedings of the Sixth Arabic Natural Language Processing Workshop*. Kyiv, Ukraine (Online): Association for Computational Linguistics, April 2021, pp. 92–104.
- [63] L. Zhang, H. Zhu, S. Brahma, and Y. Li, “Small but mighty: New benchmarks for split and rephrase,” *arXiv preprint arXiv:2009.08560*, 2020.

# Unsupervised Feature Learning Methodology for Tree based Classifier and SVM to Classify Encrypted Traffic

RAMRAJ S<sup>1</sup>, Usha G<sup>2</sup>

Research Scholar, Department of Computer Science and Engineering  
SRM Institute of Science and Technology, Kattankullathur, Chennai, India<sup>1</sup>  
Associate Professor, Department of Computing Technologies  
SRM Institute of Science and Technology, Kattankullathur, Chennai, India<sup>2</sup>

**Abstract**—Presently, sample social applications have emerged, and each one is trying to knock down the other. They expand their game by bringing novelty to the market, being ingenious and providing advanced level of security in the form of encryption. It has become significant to manage the network traffic and analyze it; hence we are performing a network traffic binary classification on one of the globally used application – WhatsApp. Also, this will be helpful to evaluate the sender-receiver system of the application alongside stipulate the properties of the network traces. By analyzing the behavior of network traces, we can scrutinize the type and nature of traffic for future maintenance of the network. In this study, we have carried out three different objectives. First, we have classified between the WhatsApp network packets and other applications using different ML classifiers, secondly, we have segmented the WhatsApp application files into image and text and third, we have incorporated a deep learning module with the same ML classifiers to understand and boost the performance of the previous experiments. Following the experiments, we have also highlighted the difference in the performance of both tree-based and vector-based classifiers of Machine Learning. Based on our findings, XGBoost classifier is a pre-eminent algorithm in the identification of WhatsApp network traces from the dataset. Whereas in the experiment of WhatsApp media segmentation, Random Forest has outperformed the other ML algorithms. Similarly, SVM when clubbed with a Deep Learning Auto encoder boosts the performance of this vector-based classifier in the binary classification task.

**Keywords**—Network traffic; encrypted network traffic; tree based classifiers; SVM

## I. INTRODUCTION

All network applications need encryption as it provides authenticity, confidentiality and integrity to the users. In unencrypted network traffic, an intruder; whether spiteful (attacker), or not (e.g. network administrator tracking infrastructure) can read network packets and can view their contents. This leads to the intrusion of privacy and misuse of user's data. Whereas, in case of WhatsApp application, the data is end-to-end encrypted from the sender to receiver. Such applications do not leave room for any kind of violation of privacy. In widespread, encryption has a giant effect on detection and analysis of network traffic, because it conceals all payload statistics. As a result, new methodologies and frameworks are required to understand the complexity of Network traces without the need for decryption.

With an efficient and accurate Network Traffic Classification (TC), we can attain the cognizance of the nature and type of packets without the need of decryption. This is a secure way for Network traffic analysers to understand about the complex features of data packet. This could benefit the Network Traffic analysers in wide area of applications including advertising, allocating more bandwidth, understanding network patterns and its alterations etc. without the need of decryption. However, the rising strength of encryption requires efficient frameworks which can sustain the complexity of different and novel features of the data packets and can yield accurate results. In this study, we have used multiple Machine learning classifiers for performing binary TC. In accordance with that, we have also incorporated Deep Learning Auto encoder and PCA with ML classifiers to see their influence on the previous results. Over the last few studies, researchers have demonstrated how the inclusion of Deep Learning in the classification frameworks has improvised the results. This extra module of DL thus takes care of the packet's features and extracts them for the classifier. With an extracted set of features, the complexity for the classifier reduces and as a result it performs better. The results in this paper indicate the same and give a clear understanding of the performance of different models taken into consideration. The model which we proposed is to make the features learned from the deep learning algorithm such as auto encoder, PCA and those features will be feeded to train the machine learning model such as SVM, XGBoost, and Random forest. The performance analysis is done to verify whether the auto encoder or PCA helps the machine learning model to improve the classification in network traffic data.

### A. Key Contributions

Our major contributions in this paper are:

- The available dataset [1] [2] does not includes WhatsApp network traffic traces. In this work we collected WhatsApp network traffic and integrate into the existing dataset.
- Implementing auto encoder, PCA with machine learning models for network traffic classification.
- Comparative study on the performance of tree based classifiers and large margin classifier for encrypted network traffic classification.

### B. Introduction of WhatsApp Application Data in the Dataset

The WhatsApp Network traces are captured using the Port Mirroring Technique in a supervised environment over a secure connection. This enables us to club this captured data with open-source datasets available and use this combined dataset to train our proposed models.

### C. Comparison of Tree-Based and Vector-Based ML Classifiers for Traffic Classification

A contrast is highlighted between Tree-based and Vector-based algorithms of Machine Learning and the results are thoroughly explained using precision and F-1 scores. At the end of this study, a claim is also made in favour of Tree-based algorithms for their excellent performance.

### D. Feature Learning using Deep Learning and Comparison of Proposed Cross-Frameworks

An extra step is implemented to distribute the workload of Traffic Classification over different modules of the proposed framework. During the feature extraction and learning, Auto encoder and PCA come into action and pass the set of learned features to the classifier. This not only performs the TC but also boosts the performance of the model.

## II. RELATED WORK

The rising demand for Network Traffic Classification (TC) [3] [4] has led to many studies in the recent years. TC has use in wide areas of applications and holds a huge demand among the Network Analysers. A lot of studies have also demonstrated hybrid models that are known to have better accuracy in identifying large variety of applications.

In [5], T.T. Nguyen and others put forth the execution of ML techniques to IP traffic classification. They claim that the algorithms have demonstrated varied accuracy, even up to 99 percent, for a wide range of web application traffic. In [6], A. Dainotti and others have provided a wide range of worthy recommendations for traffic classification. According to one of their recommendations, the blend of traffic classification and algorithms should include a thorough analysis of efficiency and performance. Weibo Liu and others in [7] bestow the combination of Auto encoder, convolutional neural networks, deep belief network, and restricted Boltzmann machine. Using this combination, they indicate that we can now use unsupervised learning algorithms to process the unlabelled data. In [1], Hongtao Shi and others propose an approach that insists on dimensional reduction in feature space and overcomes the multi-class imbalance. Giuseppe Aceto and others in [2] put forward Deep Learning to build traffic classifiers based on auto-extracted features and reflect their traffic patterns. Finally, they have dissected existing DL algorithms in standard traffic classification. In [8], Chuan Guo and others put forth the calibration prospect of the ML algorithms. Their findings signify the effectiveness of temperature scaling on datasets. Arthur Callado and others in [9] propose techniques like signature-matching, sampling, and inheritance, known in the field of IP traffic analysis, and focuses on application detection. In [10], Wei Wang and others present a new perspective of traffic classification using AI. They achieved good accuracy using a traffic classifier, which can learn features automatically

(used CNN). Meanwhile, [11] and [12] are concerned with the privacy involved in network traffic analysis in applications present on the smartphone. They propose methods to secure end-to-end encryption as well as show the threats an eavesdropper can bid.

A similar approach is also investigated in [13] where the author proposes high performance multi class classification architecture capable of enhancing the classification results by up to +9.5 percent. The popularity and efficacy of DL based hybrid model is also evident in [14]. In [14], [15], [16] a pure DL framework with a series of Neural Network is proposed. The focus here lies on addressing a novel and updated experimental setup for an umbrella of TC tasks which are encrypted. In addition to this, sustainable frameworks are designed by researchers to use it for multi-classification tasks. In [17], a single architecture is proposed which can perform two tasks simultaneously. Task one being the characterization of the network traces based on F2P and P2P. The second task being the identification of applications [18]. With this single Deep Learning framework [19], the author has been able to further distinguish the packets into VPN and Non-VPN [20] traces followed by TC. The need of standard framework in network traffic analysis is discussed in [21].

Our research is greatly influenced by the concept of hybrid models and their multi-classification purpose. With that in mind, we have deduced different models and selected the most promising among them.

In this work the feature learning process is automated through the auto encoder, PCA. The proposed model trains the machine learning model [5] with the features learned using the auto encoder, PCA. A comparative study is done between the performance of the tree based classifiers and SVM. From the results it shows the performance of random forest improved much better with auto encoder.

## III. DATASET COLLECTION

Fig. 1 demonstrates the experimental setup involved in the data collection process. This setup consists of a router, a viable internet connection, a port mirroring switch, communicating devices and software for the purpose of analysis. The Wireshark software is used on the controlling unit which displays the features and network traces in a series of timeline. However, this raw data obtained from this setup is in the .pcap format. CIC Flowmeter tool is used to convert this extension in a usable format of .csv extension for our model to be trained. Once the data is converted, it passes through the pre-processing stage. Here, the CIC Flowmeter tool extracts more features from the raw data by performing mathematical calculations using statistics at the backend. These are called the derived features. Based on the previous studies, the relevant features for our model are considered and the others are omitted. These features include both the backward BWD and forward FWD transmission flow. Once the data is pre-processed, it is combined with an open-source network traffic dataset (ISCXVPN2016) to train our proposed model. As a result, this combined dataset includes WhatsApp network traces along with Other Applications. This experimental setup is carried out in a supervised environment and is one our major contributions in this paper.

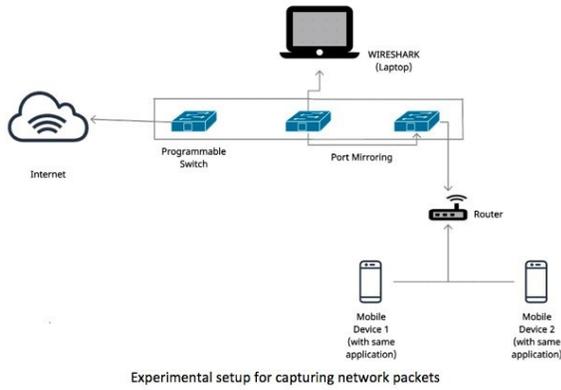


Fig. 1. Experimental setup for capturing network packets

### A. Dataset

The dataset used in the experiment includes VPN [22] traffic data from the Canada Institute for Cybersecurity. Since VPN data are encrypted it is combined with WhatsApp traffic data. Total VPN data is 16395 which includes 'vpn\_email', 'vpn\_facebook', 'vpn\_hangouts', 'vpn\_spotify', 'vpn\_youtube' and the WhatsApp data is about 17997. For experiments on classifying WhatsApp data as image, text the number of image data is 12546 and number of text data is 26258

## IV. PROPOSED METHODOLOGY

Although a simple Machine Learning Classifier is capable of distinguishing between two types of data packets and segmenting them into classes, however with the rising complexity and security of social media applications, these traditional classifiers namely SVM, Random Forest, XGBoost etc. under perform.

Our proposed model in Fig. 2 is a dual stage framework with each stage performing an independent task in the TC. This not only boosts the efficiency of the model but also yields improvised results as compared to pure Machine Learning models. Stage 1 provides a pipeline where the data is captured and pre-processed before it is passed onto the classifiers. In this stage, MinMax scaler is implemented to normalize the data entries and convert them in the range of [0,1]. ML classifiers like SVM, XGBoost and Random Forest are tested on our self-gathered dataset. Also, a contrast between Tree-based (XGBoost and Random Forest) and Vector-based (SVM) algorithms is made at the end of each stage.

In order to improve the results obtained in stage 1, stage 2 is introduced with an extra module of Deep Learning. Stage 2 provides a fusion of Machine Learning and Deep Learning to form a Hybrid System. This system comes into action after the data passes through the pre-processing pipeline. For the purpose of comparison, Auto encoders and PCA are used with each classifier implemented in stage 1. Once the data is passed through the Auto encoders or PCA, it extracts the complex features from the dataset and provides a reduced set of relevant features which are then traversed back to stage 1. Here the normal flow of data is then followed by Machine Learning.

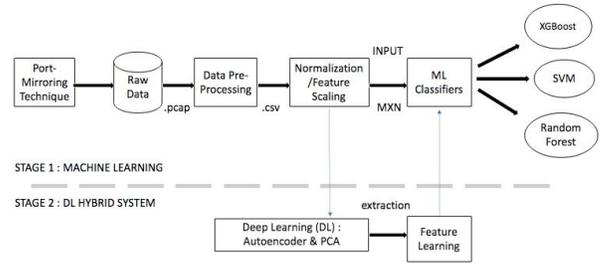


Fig. 2. A dual-stage hybrid architecture

This simple yet effective system has proved to enhance the outcome of classification obtained from stage 1.

The Random Forest algorithm is a tree-based algorithm suitable for selecting relevant features to perform classification. In the above algorithm, the tree starts at a single root node. At each node of the tree, a subset from the feature set is taken which is then split into different nodes. Each node denotes an attribute relevant for the classification of the Network Packets.  $f$  denotes the subset of features taken for each node, where  $f$  is much smaller than  $F$ . The decision to split a particular node is a computationally expensive process. By restricting each split in the tree, the rate of learning becomes faster.

### Algorithm 1 Algorithm for Feature Learning with Auto Encoder

**Input:**[X1,X2,X3,...]

- 1: Feature in  $X1=X1,X2,...,Xn$
- 2: **for each** X **do**
- 3:      $y=f(x)=se(w*x+bh)$
- 4:      $se(x) = \text{sigmoid}(x)=1 / (1+e^{-x})$
- 5:      $g(y)=sd(w*y+bf)$
- 6:      $sd = \text{tanh}(x)(e^x-e^{-x}) / (e^x+e^{-x})$
- 7:     Optimize  $\theta=[W,bh,b'r]$
- 8:     **while**  $D_i = [x_1, x_2, x_3, \dots, x_n]$  **do**
- 9:          $JAE(\theta)=\sum R(x,r) \quad x \in D_i$
- 10:     **end while**
- 11: **end for**
- 12: Train\_Feature =
- 13: Test\_Feature=
- 14: **for each** Train\_X in  $Se(X)$  **do**
- 15:     Train\_Feature = X
- 16: **end for**
- 17: **for each** Test\_X in  $Se(X)$  **do**
- 18:     Test\_Feature = X
- 19: **end for**
- 20: se – Encoder Function
- 21: sd – Decoder Function

In Algorithm 1, where  $R$  denotes the reconstruction error,  $w'$  the weights given to the inputs of the hidden layer, and  $b'$  the biasness of the inputs given to the hidden layer.  $Train_X$ ,  $Test_X$  denotes the number of training and testing samples. The  $Train_{Feature}$ ,  $Test_{Feature}$  are used to train the machine learning models. Encoder encodes the input  $X_i$  to hidden representation  $h_i$ . It does with the function  $h(X)=G(W*X+B)$ .  $W$  is the set of weights,  $B$  is bias and  $G()$  is a nonlinear

function. The work of the decoder is to reconstruct the input from the hidden representation. Initially the weights and bias are randomly assigned and the values are optimized with every iteration. The loss function is used to calculate how much the hidden values are deviated from the original data. The network traffic data with dimensions of 21 features are given as input to the encoder. The encoder with sigmoid activation function produces a hidden representation of data with 10 dimensions. The loss function used for the decoder is binary cross entropy. The Adam optimizer is used for getting the right set of values for W and B. After the encoder optimizes the values, the features are transformed to train the machine learning models. The machine learning model such as SVM, XGBoost, and Random Forest are trained and tested with the features from auto encoder, PCA.

## V. RESULTS AND DISCUSSION

TABLE I. COMPARISON OF ML CLASSIFIERS FOR WHATSAPP AND OTHER APPLICATIONS

Model	Precision	Recall	F1 Score
Random Forest	0.84	0.99	0.91
XGBoost	0.98	0.97	0.98
SVM	1.00	0.90	0.94

In this experiment, we aim to classify the WhatsApp network packets from other applications. Evaluation of this experiment is conducted using a self-gathered dataset. The dataset consists of encrypted traffic [23] flow which is gathered using a port-mirroring switch and a network router. The raw data consists of more than 35 features which are eventually narrowed down to 21 features after doing a considerable amount of data analysis. The influence of original (non-normalized) and normalized data has been studied in this experiment. The encrypted data is then normalized using MinMax Scaler for our distance based algorithms to yield correct and accurate results.

Towards the end of this experiment, we propose a comparison between vector based and tree-based machine learning classifiers. The normalized traffic flow data is fetched into three machine learning algorithms, namely, SVM, XGBoost and Random Forest algorithm. The comparison between these three classifiers is illustrated in Table I. It demonstrates the effectiveness and performance of tree-based algorithms over the vector-based classifiers. With a F-1 score of 0.92 (for WhatsApp) and 0.98 (for others), XGBoost succeeds the other classifiers in this experiment. With the analysis of the F-1 scores derived above, we conclude that tree-based algorithms perform better than vector-based algorithms in the classification of network packets.

In the second experiment we have presented the comparison between different machine learning classifiers in the classification of image and text files of WhatsApp Network packets. The network packets obtained using the Port mirroring switch consists of a combination of different file transmissions. Due to their encrypted nature, it becomes challenging to classify them into different classes based on the nature of these files. As a result, tree- based and vector-based classifiers in machine learning are used for this purpose. To have a better

understanding, the author has denoted the media files as class 1 and text files as class 0. For each class, precision, recall and F-1 score is calculated to measure and analyze the performance of these classifiers. Table II gives the performance metrics for each class and its classifier. In Table II, Random Forest and XGBoost are the tree-based classifier which follows the approach of branching for classification tasks. On the other hand, SVM follows a vector- based approach to classify the packets. Upon comparison, it is found that Random Forest overcomes the performance of XGBoost and SVM. With an F-1 score of 0.91 (class 0) and 0.80 (class 1) for image files and text files classification, Random Forest has been the best classifier among the other two algorithms. Followed by Random Forest, it is noticed that XGBoost is closer to Random Forest in terms of performance with an F-1 score of 0.86 (class 0) and 0.67 (class 1). From this experiment, we conclude that tree-based algorithms perform better than vector-based classifiers and Random Forest has achieved better results than XGBoost and SVM in the classification of image and text files of WhatsApp packets.

TABLE II. CLASSIFICATION OF WHATSAPP FILES INTO TEXT AND IMAGE USING ML CLASSIFIERS

Model	Precision	Recall	F1 Score
Random Forest	0.88	0.95	0.91
XGBoost	0.83	0.90	0.86
SVM	0.72	0.91	0.80

TABLE III. COMPARISON OF PROPOSED HYBRID SYSTEMS FOR WHATSAPP CONTENT CLASSIFICATION

Model	Precision	Recall	F1	Accuracy
Auto Encoder + SVM	0.80	0.80	0.84	0.90
Auto Encoder +XGBoost	0.83	0.88	0.85	0.91
Auto Encoder +Random Forest	0.85	0.84	0.84	0.88
PCA + SVM	0.68	0.98	0.80	0.85
PCA + XGBoost	0.73	0.66	0.69	0.78
PCA + Random Forest	0.70	0.95	0.81	0.84

In the last experiment, we present a method and a framework for efficient and effective feature extraction followed by ML classification. This architecture consists of deep learning modules and machine learning classifiers in order to fulfill the objective of our experiment i.e. classification of WhatsApp network packets into text and image. During this experiment, we show that our implementation of the framework can extract the features of the network packets which are encrypted and unlabelled. For the purpose of comparison, Auto encoders and PCA are taken into consideration for feature extraction.

As we move forward in this architecture, tree-based and vector-based machine learning classifiers are implemented to classify the packets based on the features extracted. For each feature extraction module, three classifiers are tested, namely, SVM, XGBoost and Random Forest. Table III highlights the results obtained after testing all the algorithms. SVM when implemented alone performed the lowest among XGBoost and Random Forest as observed in experiment 2. However, in contrast to this, it is seen that auto encoders are improvising the results of SVM in experiment 3. With the use of Auto encoders, the F-1 scores of SVM have drastically improved

whereas it has been nearly same for the other two classifiers. From the chart shown in Fig. 3 and 4, it is clear that PCA has failed to show any improvement in the classification task, the results suggest that feature extraction through auto encoders has contributed towards a positive learning curve. In conclusion, the deep learning module – auto encoders has provided a better result when clubbed with SVM in comparison to when SVM is implemented alone. Fig. 5 and 6 shows how the precision and recall value varies between the different models with respect to auto encoder, PCA usage.

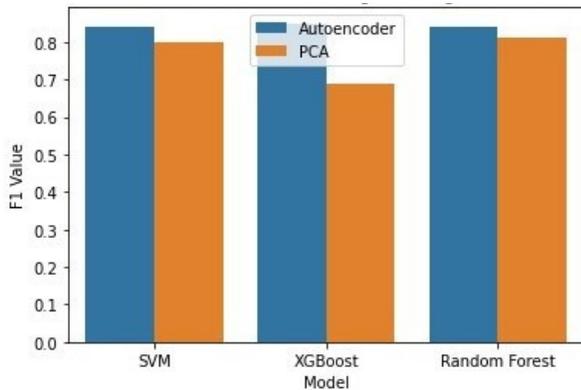


Fig. 3. Comparison of auto encoder and PCA with machine learning model in classifying WhatsApp image from text

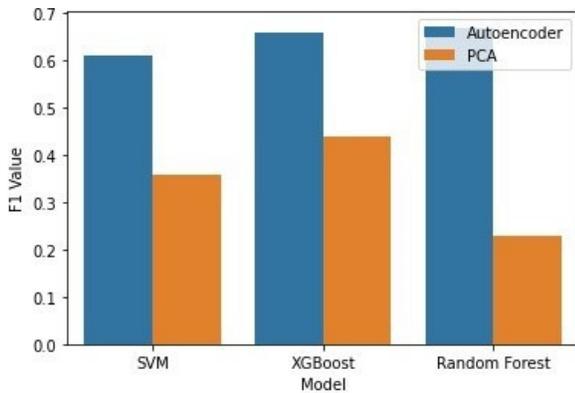


Fig. 4. Comparison of auto encoder and PCA with machine learning model in classifying WhatsApp text from image

## VI. CONCLUSION AND FUTURE WORK

Every tactical model performed differently under our experimental environment. This is also attributed to the dataset as well as the computational complexity of the algorithms. Thus, the model yielding the best results should also be efficient enough to perform TC on large datasets. It should also be noted that the efficiency of the model is directly related to its performance in real time TC. The computational complexity of each algorithm is closely scrutinized. The vector-based classifier SVM has a computational complexity of  $O(n^3)$ , wherein  $n$  is the training data's strength. While the computational complexity of Tree based algorithms highly depends upon the number of attributes taken into consideration which is 45

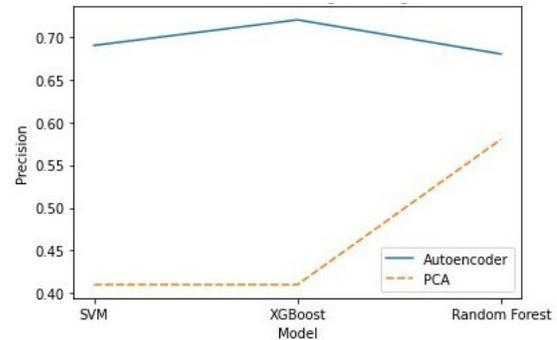


Fig. 5. Comparison of auto encoder and PCA with machine learning model in classifying WhatsApp image from text in terms of precision

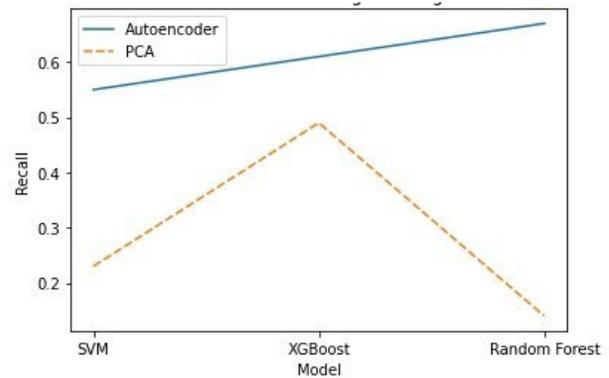


Fig. 6. Comparison of auto encoder and PCA with machine learning model in classifying WhatsApp image from text in terms of recall

features in our case. Thus, the feature count becomes directly proportional to the number of subtrees formed in the model. As a result, the computational complexity of XGBoost for learning each tree becomes  $O(n \log n)$ . In the case of Random Forest algorithm, the computational complexity is  $O(TD)$ , where  $T$  is the size of random forest and  $D$  is the maximum depth. In case of Random Forest, the subtree balance and  $D$  highly influence the results. Different tactical models and hybrid systems are tested in our paper and a conclusion is drawn in favour of the Tree-based classifiers. Upon testing all the classifiers upon our self-gathered dataset, it is concluded that Tree-based classifiers (XGBoost and Random Forest) outperforms the Vector-based classifier (SVM) and yields a better accuracy and F-1 score in Network TC. Therefore, from experiment 1, Performance (XGBoost) is greater than the Performance of Random Forest and SVM. Similarly, in experiment 2, the performance of Random Forest and XGBoost are interchanged whereas SVM remains the last in comparison. This clearly indicates that Tree-based classifiers are better in performance than Vector-based classifiers. Also, the use of Deep Learning for feature extraction has given a boost to the results of SVM. Thus, auto encoder reduces the complexity of the features and supports the classifiers in the classification process.

Certain cases are expected to be covered in the future work for making this proposed architecture a state-of-the-art system. This includes the segmentation of other media files including file transfer, voice message and location sharing.

Apart from this, other Deep Learning models like CNN, Deep Neural Networks, RNN, etc. are yet to be tested upon this dataset. Thus, a further investigation is required with other Deep Learning modules. In future the work can be extended to other WhatsApp data such as text, voice, etc. The development of an intrusion detection system for encrypted data such as WhatsApp, Telegram.

#### REFERENCES

- [1] H. Shi, H. Li, D. Zhang, C. Cheng, and X. Cao, "An efficient feature generation approach based on deep learning and feature selection techniques for traffic classification," *Computer Networks*, vol. 132, pp. 81–98, 2018.
- [2] G. Aceto, D. Ciunzo, A. Montieri, and A. Pescapé, "Mobile encrypted traffic classification using deep learning," in *2018 Network traffic measurement and analysis conference (TMA)*. IEEE, 2018, pp. 1–8.
- [3] H. Wang, S. Zhou, H. Li, J. Hu, X. Du, J. Zhou, Y. He, F. Fu, and H. Yang, "Deep learning network intrusion detection based on network traffic," in *International Conference on Artificial Intelligence and Security*. Springer, 2022, pp. 194–207.
- [4] M. H. Rahman, R. B. Mofidul, and Y. M. Jang, "Spectrum based wireless radio traffic classification using hybrid deep neural network," in *2022 Thirteenth International Conference on Ubiquitous and Future Networks (ICUFN)*. IEEE, 2022, pp. 95–99.
- [5] T. T. Nguyen and G. Armitage, "A survey of techniques for internet traffic classification using machine learning," *IEEE communications surveys & tutorials*, vol. 10, no. 4, pp. 56–76, 2008.
- [6] A. Dainotti, A. Pescapé, and K. C. Claffy, "Issues and future directions in traffic classification," *IEEE network*, vol. 26, no. 1, pp. 35–40, 2012.
- [7] W. Liu, Z. Wang, X. Liu, N. Zeng, Y. Liu, and F. E. Alsaadi, "A survey of deep neural network architectures and their applications," *Neurocomputing*, vol. 234, pp. 11–26, 2017.
- [8] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *International conference on machine learning*. PMLR, 2017, pp. 1321–1330.
- [9] A. Callado, C. Kamienski, G. Szabó, B. P. Gero, J. Kelner, S. Fernandes, and D. Sadok, "A survey on internet traffic identification," *IEEE communications surveys & tutorials*, vol. 11, no. 3, pp. 37–52, 2009.
- [10] W. Wang, M. Zhu, X. Zeng, X. Ye, and Y. Sheng, "Malware traffic classification using convolutional neural network for representation learning," in *2017 International conference on information networking (ICOIN)*. IEEE, 2017, pp. 712–717.
- [11] V. F. Taylor, R. Spolaor, M. Conti, and I. Martinovic, "Robust smartphone app identification via encrypted network traffic analysis," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 1, pp. 63–78, 2017.
- [12] B. Saltaformaggio, H. Choi, K. Johnson, Y. Kwon, Q. Zhang, X. Zhang, D. Xu, and J. Qian, "Eavesdropping on {Fine-Grained} user activities within smartphone apps over encrypted network traffic," in *10th USENIX Workshop on Offensive Technologies (WOOT 16)*, 2016.
- [13] R. Schuster, V. Shmatikov, and E. Tromer, "Beauty and the burst: Remote identification of encrypted video streams," in *26th USENIX Security Symposium (USENIX Security 17)*, 2017, pp. 1357–1374.
- [14] P. Velan, M. Čermák, P. Čeleda, and M. Drašar, "A survey of methods for encrypted traffic classification and analysis," *International Journal of Network Management*, vol. 25, no. 5, pp. 355–374, 2015.
- [15] R. Chapaneri and S. Shah, "Enhanced detection of imbalanced malicious network traffic with regularized generative adversarial networks," *Journal of Network and Computer Applications*, vol. 202, p. 103368, 2022.
- [16] M. Lotfollahi, M. Jafari Siavoshani, R. Shirali Hossein Zade, and M. Saberian, "Deep packet: A novel approach for encrypted traffic classification using deep learning," *Soft Computing*, vol. 24, no. 3, pp. 1999–2012, 2020.
- [17] A. Rago, G. Piro, G. Boggia, and P. Dini, "Multi-task learning at the mobile edge: An effective way to combine traffic classification and prediction," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 9, pp. 10 362–10 374, 2020.
- [18] S. Dong, "Online encrypted skype identification based on an updating mechanism," *arXiv preprint arXiv:2203.12141*, 2022.
- [19] S. Rezaei and X. Liu, "Deep learning for encrypted traffic classification: An overview," *IEEE communications magazine*, vol. 57, no. 5, pp. 76–81, 2019.
- [20] L. Chen, Y. Xue, Y. Mu, L. Zeng, F. Rezaeibagha, and R. Deng, "Case-sse: Context-aware semantically extensible searchable symmetric encryption for encrypted cloud data," *IEEE Transactions on Services Computing*, 2022.
- [21] J. Holland, P. Schmitt, P. Mittal, and N. Feamster, "Towards reproducible network traffic analysis," *arXiv preprint arXiv:2203.12410*, 2022.
- [22] "VPN-nonVPN dataset (ISCXVPN2016)," [21] <https://www.unb.ca/cic/datasets/vpn.html>, accessed: 2022-08-30.
- [23] D. F. Isingizwe, M. Wang, W. Liu, D. Wang, T. Wu, and J. Li, "Analyzing learning-based encrypted malware traffic classification with automl," in *2021 IEEE 21st International Conference on Communication Technology (ICCT)*. IEEE, 2021, pp. 313–322.

# Indoor Pollutant Classification Modeling using Relevant Sensors under Thermodynamic Conditions with Multilayer Perceptron Hyperparameter Tuning

Percival J. Forcadilla

Department of Computer Science, College of Science  
University of the Philippines Cebu  
Lahug, Cebu City, Philippines

**Abstract**—Air pollutants that are generated from indoor sources such as cigarettes, cleaning products, air fresheners, etc. impact human health. These sources are usually safe but exposure beyond the recommended standards could be hazardous to health. Due to this fact, people started to use technology to monitor indoor air quality (IAQ) but have no capability of recognizing pollutant sources. This research is an improvement in building a classification model for recognizing pollutant sources using the multilayer perceptron. The current research model receives four data parameters under warm & humid and cool & dry conditions compared to nine parameters of the previous literature in detecting five pollutant sources. The classification model was optimized using GridSearchCV to obtain the best combination of hyperparameters while giving the best-fit model accuracy, loss, and computational time. The tuned classification model gives an accuracy of 98.9% and a loss function value of 0.0986 under the number of epochs equal to 50. In comparison with the previous research, the accuracy was 100% with the number of epochs equal to 1000. Computational time was greatly reduced at the same time giving the best-fit accuracy and loss function values without incurring the problem of overfitting.

**Keywords**—Indoor air pollutants; pollutant sources; indoor air quality; IAQ; sensors; multilayer perceptron; classification modeling; gridSearchCV; hyperparameter tuning

## I. INTRODUCTION

In the last several years, scientific evidence has been produced regarding indoor air being more harmful than the air outdoors. That is why, people at homes, offices, and schools desire to have indoor air to be fresh and pleasant, not harmful to health to promote productivity. Air pollutants are one of the contributory factors in harming indoor air which comes from different sources and significantly impacts the health of the people in the long run. With this threat present, relevant organizations develop Indoor Air Quality (IAQ) standards and guidelines to eliminate or prevent people from the potential health risk of these indoor pollutants. These organizations gave threshold values for concentration levels and exposure time in maintaining the optimal IAQ level.

Aside from the IAQ standards and guidelines made by different organizations around the world, the research community contributed a lot to the development of advanced technologies that will help monitor and keep IAQ at the optimum level. Research themes like performance assessments on industry and consumer sensors and devices, factors affecting indoor air pollution in residential and commercial spaces,

machine learning algorithms, strategies, and techniques were also included giving new contributions to the existing body of knowledge. But with all these research trends, it was found that limited contribution was given to IAQ research on pattern recognition techniques. This gave the researcher motivation to further study the topic revolving around machine learning using pattern recognition techniques.

One of the studies that were considered was from Saad Shaharil Mad et al. [19]. In the study, nine parameters were used to develop pollutant recognition based on pattern recognition techniques. With the motivation of exploring the topic of pattern recognition techniques, the researcher formulated a research question to further explore the topic. The question was - How effective is an algorithm if the dataset parameters are reduced from nine to targeted parameters in classifying indoor air pollution sources in terms of model accuracy and loss? The research question was then supported with another study which gives this research a starting point for further investigation. That study was from Demanega Ingrid et al. [19] and was all about assessing the performance of low-cost monitors and not classifying indoor pollutants. The result of the study shows positive detection outcomes for all sensors and devices using particulate matter (PM), carbon dioxide (CO<sub>2</sub>), total volatile organic compounds (TVOC), indoor air temperature, and relative humidity as the parametric specifications considered in the research. Most of the low-cost sensors in the study responded to the simulated indoor pollutant sources with varying parametric readings. Also, the pollutant sources in the study of Demanega Ingrid et al. [11] and Saad Shaharil Mad et al. [19] shows to have relatively the same sources used as indoor pollutants. On that note, a hypothesis was then created to predict that selecting key parameters out of the nine parameters that were considered in the study of Saad Shaharil Mad et al. [19] can still be able to recognize or classify indoor air pollutants. With this study, the sensors to use were purely dependent on consumer-grade monitors available in the market which has no ability to recognize sources. The selected parameter that was considered were PM, CO<sub>2</sub>, TVOC, and formaldehyde (HCHO). These parameters were collected in a controlled room setup where indoor air temperature and relative humidity were set as one parameter while generating an environment simulation commonly encountered in many indoor climates around the world. These indoor climates are set as two thermodynamic conditions, warm & humid and cool & dry conditions which were adopted in the research

of Demanega Ingrid et al. [11]. The multilayer perceptron classification model algorithm was used as the only algorithm for this study since it performs to have the highest classification accuracy in the previous study. Lastly, to further improve the previous work and set an additional contribution, optimization was done through hyperparameter tuning.

This paper starts by introducing what Indoor Air Quality is followed by the significance of the study, and the different research trends that bring motivation to implementing this research. Validation of the research question was done through the review of the literature followed by a brief description of the methodology. The methodology starts with data collection of the target parameters under two thermodynamic conditions, processing the data through a pattern recognition technique (Multilayer Perceptron) leading to the classification of pollutant sources. Afterward, a preliminary analysis was done to show the parameter's categorical correlation followed by the model's accuracy and loss results. Finally, an optimization process through hyperparameter tuning was also performed.

## II. LITERATURE REVIEW

### A. Indoor Air Quality: Description, Sources and Effects to Human Health

The quality of air inside a building or structure like homes, schools, and offices which promotes good health and comfort for occupants is called indoor air quality (IAQ). It is also the extent to which human requirements in terms of indoor air are met. In this case, it is the desire which air should be fresh and pleasant, have no harmful effect on health, promotes good working conditions in the context of building offices or homes, and productivity in terms of learning at schools [1]. IAQ has been recognized as an important factor in people's health and comfort in indoor environments because 90% of people's time is spent indoors [2]. Also, studies have shown that the occupants are 100 times more exposed to indoor air pollutants than outdoor air pollutants and the concentration of air pollution indoor is seen to be 2 to 4 times higher than that of outdoor [3].

When indoor air quality is not taken into consideration for buildings and structures, possible health concerns due to high indoor pollutant exposure may be experienced. According to United State Environmental Protection Agency (EPA), there are many sources of indoor air pollution. Sources like fuel-burning combustion appliances, tobacco products, products for home cleaning and maintenance, building materials, furnishings, a product like air fresheners, and many more [4]. Over the years, these sources have been producing gaseous pollutants which are chemicals tagged as hazardous. These include radon, ozone, nitric oxides, sulfur dioxide, carbon monoxide, diatomic carbon, and VOCs [5]. Aside from these sources, particulate matter from combustion and cleaning activities, heavy metals from fuel consumption and building materials, airborne particles, pest control chemicals, and biological contaminants are recognized as air pollutants harmful to human health [6].

With the threat of pollutant exposure leading to significant risks to human health, relevant organizations developed different IAQ standards and guidelines. Among these organizations are the World Health Organization (WHO) and the United

States Environmental Protection Agency (USEPA) which contributed to the making of indoor air quality standards and guidelines. The guidelines serve as a database reference to eliminate or prevent people from too much exposure to indoor air pollution and the potential health risks that may be brought to the human population [7]. Aside from WHO and USEPA, other recognized organizations like ASHRAE [22], National Health & Medical Research Council in Australia, and many more around the world have set guidelines and threshold values to maintain an optimal IAQ [8].

### B. IAQ Current Research Trends

With the standards and guidelines set by the different organizations and the development of sensor technologies in a network and systems, different research trends have been developed through time. Most of this literature fall under advanced technologies for monitoring, performance assessments of cheap and high-end monitoring device or systems, factors affecting indoor air pollution in residential and commercial spaces, forecasting, and pattern recognition techniques.

1) *Advanced technologies for monitoring IAQ:* The development of mobile technologies and the internet of things (IoT) has brought great capability toward improving IAQ monitoring systems. Air quality monitoring nowadays can easily be done through smartphones by simply accessing the web. In the study of M. Tastan et al. the research proposed an IoT-based real-time e-nose measuring system using low-cost electronic sensors [9]. The system includes sensors such as MH-Z14A for Carbon Dioxide, MICS-4514 for Nitrogen Oxide and Carbon Monoxide sensor, GP2Y1010AU as a dust sensor, and DHT22 for Temperature and Humidity sensor together with an ESP32 microcontroller with built-in Wi-fi, used to process the information provided by the sensors. Pollutant concentration data is thrown into the Blynk cloud, and an android-based mobile user interface was developed for users to access the data in a form of digital or graphical monitoring.

In another study by Giacomo Chiesa et al. [10], a system was developed based on several sensor data to model the IAQ which serves as input in controlling the ventilation system. The system is connected to an app that features management of the device, real-time data visualization, and statistical data [10]. Device management allows the user to create a list of installed devices and set desired ventilation time to report indoor conditions in threshold value for each device. Real-time data visualization includes the quality level of indoor air and each of the parameters because of the sensor devices, and lastly, statistical data which gives users daily or weekly graphs for each sensor. The parameters that were considered for this research were carbon dioxide, TVOC, pressure, humidity, and temperature. Raspberry Pi serves as the backend platform where it handles device management, sensor-to-microcontroller interface, sensor data communication, control algorithms, data storage, and aggregation. The research data platform used for IoT needs was MongoDB. It is a source-available cross-platform document-oriented database program.

2) *Performance assessment on IAQ monitoring device and systems:* Another research trend from the existing body of knowledge revolves around an IAQ theme that was based on IAQ monitoring devices and system performance assessment

and evaluation. Performance assessment is needed since there are different sensor devices and systems in the market and these sensors can either be low-cost or industry-grade sensors that influence the sensor reading's accuracy. Sensor devices and systems were tested to know their performance reliability or prove if these devices can competently measure a target parameter. In the study of Ingrid Demanega et al. consumer environmental monitors available in the market together with low-cost single variable sensors were assessed to know their reading reliability in thermodynamic conditions [11]. Different sources were introduced inside a room chamber like candle burning, mosquito coil burning, wood lacquer drying, room deodorant injection, essential oil heating, carpet vacuuming, popcorn cooking, and carbon dioxide injection. The parameters to be evaluated were the particulate matter (1.0, 2.5, and 10), carbon dioxide, and TVOC. Particulate Matter Monitor miniWRAS, LI-COR 850 Bioscience gas analyzer for carbon dioxide, and GrayWolf AdvancedSense Pro as TVOC monitor were all used as the industry-grade reference monitor to evaluate the consumer-grade monitors and sensors.

In another research, Zhiqiang Wang et al. [12] tested the performance of low-cost IAQ monitors for PM 2.5 and 10 [12]. The low-cost monitors are Air Quality Egg 2018 version, IQAir Airvisual Pro, Awair 2nd Edition, Kaiterra Laser Egg 2, PurpleAir Indoor, and Ikaair with reference measurement systems together with professional-grade particulate monitors. The test chamber used in this research is a room of 120m<sup>3</sup> with three external walls, two doors, and raised ceiling. Several sources were used in multiple experiments, sometimes with variations that include measuring indoor concentrations of infiltrated outdoor PM<sub>2.5</sub> were collected to evaluate the performance of the low-cost devices.

3) *Factors influencing indoor air pollution:* There are different factors that contribute to the level of indoor air pollution inside buildings and structures. Several researchers have spent time and effort discovering such factors with the aim to improve air quality inside buildings and structures. In the study of Wonho Yang et al. [13], the IAQ level was investigated in different schools in Korea with the intention to know the correlation with respect to the age of the buildings [13]. Air samples were taken indoors and outdoors with parameters to consider such as carbon monoxide, carbon dioxide, particulate matter (PM<sub>10</sub>), total microbial count, total volatile organic compounds, and formaldehyde. Results show that the factors which contributed to indoor air pollution in schools are unsatisfactory ventilation and chemical emissions from building materials or furnishings. Significant high concentrations of carbon dioxide, TVOCs, and HCHO were also found at schools constructed within one year.

A relevant factor that influences the variation of indoor pollutants inside the buildings and structures points to the different seasons. In the study of Corinne Mandin et al. [14], a European project OFFICAIR was made to gain more knowledge with regard to IAQ in modern office buildings. It was found that significantly higher concentrations of formaldehyde and ozone were measured in summer whereas benzene,  $\alpha$ -pinene, D-limonene, and nitrogen dioxide were significantly high during winter. Another study focused on the variation in the concentration of pollutants at different locations in India. The study of Arindam Datta et al. [15] focuses on

the indoor air quality of non-residential buildings which is scarce in India [15]. It was found that among different non-residential buildings, a lower concentration of pollutants was recorded in the educational building compared to the two office buildings. A ductless air-conditioning system with poor air circulation and active air filtration contributed to the higher concentration of PM<sub>2.5</sub>. In Doha, Qatar, another study was carried out to investigate the level of PM<sub>2.5</sub> and PM<sub>10</sub> in office environments [16]. It was found that the cause for significant concentrations of particulate matter inside the office is ventilation, faulty envelopes, and windows.

Different factors influencing indoor air quality have been a trend in the research community. Aside from the identified factors and themes that were done by the researchers above, the study of Mehzabenn Mannan et al. [8] summarizes some of the factors that were gathered by the researcher's related literature which influenced indoor air pollution. The research identified factors like different indoor building materials and few surface finishes and appliances, nearby construction activity, indoor movement, tobacco smoke, and computer operation, high benzene concentration was observed in the lower-level classrooms and school carpet was seen to be responsible for higher PM. Another factor indicated to be the source of air pollution was the concrete additive in an office while comparing two office settings (Beijing and Stockholm) and the contribution of indoor air pollution with respect to newly built and refurbished office buildings. Factors like occupant behavior, the role of humans, respiratory emissions from human beings, and the reaction of ozone to skin lipids are seen to be giving a contribution to indoor air pollution.

4) *Machine learning and statistical modeling in IAQ:* Building a system that can forecast the concentration level of pollutants to characterize indoor air quality has been a long important topic in the community and domain of indoor environment and health science [17]. In real-occupied environments, statistical modeling has great potential to explore and predict indoor air pollution concentration levels [18]. Statistical modeling on IAQ can use forecasting techniques to predict the level of IAQ and pattern recognition techniques which allow the system to recognize certain types of smell [19].

In the study of Wenjuan Wei et al. [18], a summary of common machine learning and statistical modeling methods was collected through a literature review. Methods were compared based on strengths and weaknesses while discussing how and where the methods have been used in the field of IAQ. A summary of machine learning algorithms used in IAQ are based on either supervised or unsupervised learning with a different type of response variable and linearity of the model.

a) *Forecasting techniques:* A study about forecasting indoor concentration levels in an office space using machine learning was made by Johanna Kallio et al. [20]. The research contribution gives the body of knowledge a comprehensive dataset covering a full year with the applicability of four machine learning methods which include ridge regression, decision tree, random forest, and multilayer perceptron. Accuracy was evaluated with respect to the different methods of prediction, history window time frames, and the impact of multiple sensor modalities. In another study, Shisheng Chen et al. [23] use a machine learning approach to predict CO<sub>2</sub>, TVOC, and HCHO. Five Classrooms at the National University

of Singapore were used to perform the continuous parametric collection of data. The dataset was trained and tested using Support Vector Machine, Gaussian Processes, M5P, and a backpropagation neural network. According to Wenjuan Wei et al. [18], an artificial neural network (ANN) is the most popular method for the prediction of IAQ based on the researcher's findings. Dwellings, offices, schools, and subway stations are the different sites where ANN modeling was used to predict several IAQ parameters [18].

*b) IAQ pattern recognition techniques:* Forecasting techniques prove to be relevant and saturated in the field of indoor air quality. In the case of pattern recognition techniques, little literature has been found to have a relevant contribution to the IAQ research community. One of which was a study that uses pattern recognition techniques to recognize specific types of pollutants by Saad Shaharil Mad et al. The authors of this topic publish two papers in 2015 and 2017. The first paper classifies pollutant sources with the use of one pattern recognition technique, ANN. The other study was an enhancement to the previous paper which utilizes different supervised machine learning algorithms like multilayer perceptron, k-nearest neighbors, and linear discrimination analysis. Nine (9) parameters were used to classify five (5) different pollutant sources. These parameters were nitrogen oxide, carbon dioxide, ozone, carbon monoxide, oxygen, VOCs, particulate matter, temperature, and humidity [19].

### C. IAQ Parameter and Sensors

The research topic regarding indoor air quality was narrowed down to the context of classifying pollutant sources using pattern recognition techniques. To classify indoor air pollutant sources, sensors must collect different parametric data to generate the IAQ dataset. Choosing the target parameters with the corresponding sensors serves as one of the important points to make this research a success.

*1) Relevant IAQ parameter consideration:* In the study of Saad Shaharil Mad et al. [19], sensors were used to target the different parameters like CDM4161 for carbon dioxide, TGS5342 for carbon monoxide, TGS2602 for VOC, MiCS2610 for Ozone, MiCS2710 for nitrogen dioxide, KE25 for oxygen, HSM20G for Temperature and Humidity, and GP2Y1010AUOF for Particulate Matter (PM10). These sensors were used to respond to the five indoor pollution sources, such as the ambient air, combustion activity, chemical presence, fragrance product, and food & beverages [19].

In another study, H. Zhang et al. [21] developed a low-cost IAQ multi-pollutant monitoring system using raspberry pi. Different sensors with specifications from different manufacturers together with their prices were carefully considered to be included in the Low-Cost Air Quality System or LCAQS. Sensors that measure Relative Humidity, Temperature, Particulate Matter (PM<sub>2.5/10</sub>), Nitrogen Dioxide, Sulfur Dioxide, Carbon Dioxide, Carbon Monoxide, Ozone, and Total Volatile Organic Compound (TVOCs) was used to develop the system [21].

To measure the indoor air quality level, parameters were divided into categories: physical condition, chemical contaminants, biological contaminants, and other common IAQ parameters. Using these categories, the study of Saad Shaharil

Mad et al. divided the sensors into three types: gas sensor, particle sensor, and thermal sensor [19]. The two studies above both have common thermal and particle parameters (Temperature, Relative Humidity, and Particulate Matter) with different IAQ research themes. Both studies above utilize nine (9) parameters, and most of the gas parameters were common to both studies except for oxygen and sulfur dioxide. In the study of Demanega Ingrid et al. [11], four parameters were only utilized to assess low-cost environmental monitors and single sensors which were used to respond to different indoor pollution sources. These parameters were temperature, relative humidity, carbon dioxide, and particulate matter [11].

Also, it can be seen in the study of Demanega Ingrid et al. [11] that the results of using the different simulated activities for indoor pollution sources show positive detection outcomes for all sensors and devices used in the study. Different particle sensor responds to indoor pollution sources such as candle burning, mosquito oil burning, and popcorn cooking. This reflects that the current study has the basis to use only target parameters to classify pollutant sources.

*2) Thermodynamic conditions:* To simulate an environment that is commonly encountered indoors in many climates around the world, thermodynamic conditions should be considered which helps assess the performance based on the standard thermal comfort zone [11,22]. This methodology was also adopted in this research in using thermodynamic conditions – warm & humid (26 +/- 1 C, 70 +/- 5%) and cool & dry (20 +/- 1 C, 30 +/- 5%)[11].

### D. IAQ Research Motivation and Summary

The exhaustive search done on the internet through online research databases proves that the topic of indoor air quality revolves mostly around advanced technologies for monitoring, performance assessments of cheap and high-end monitoring devices or systems, factors affecting indoor air pollution in residential and commercial spaces, and research about machine learning specifically forecasting techniques. On that note, a limited contribution was found regarding pattern recognition techniques in IAQ research, thus, this research takes that route related to pattern recognition. One of the pattern recognition studies that were reviewed came from Saad Shaharil Mad et al. which became the major motivation for doing the current research. The research done by Saad Shaharil Mad et al. considers nine parameters to classify pollutant sources. This pique the interest of the researcher to formulate a research question in optimizing the previous strategy. A strategy to select a few parameters out of the nine, to classify pollutant sources. In selecting parameters to be included in the classification process, the study of Demanega et al. shows results that were relevant in choosing the parameter of this research. To add another layer of parameter, the thermodynamic condition was also adopted to simulate two indoor climates commonly encountered.

## III. METHODOLOGY

This chapter discusses the key design choices, concepts, and procedures in attaining the classification of Indoor Air Pollutant Sources using targeted Pollutant Parameters based on Machine Learning's Pattern Recognition Techniques for Indoor

Air Quality (IAQ) Systems. Fig. 1 below shows the outline of the methodological process on how to attain the research objectives.

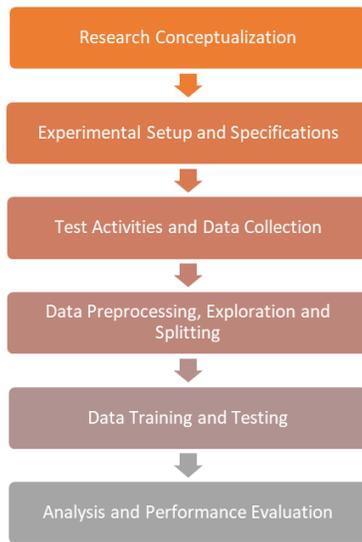


Fig. 1. Research framework flowchart

After establishing the research novelty, relevance, and feasibility through the review of related literature, the development procedure of this research starts with the implementation stage through research conceptualization. This defines the research considerations regarding the approach, initial strategies, the participants involved, and the research setting. Experimental setup and specifications were the next procedure which provides details regarding the room setup, thermodynamic conditions, sources of indoor air pollutants to consider, IAQ parameters, and what device monitor to use. When the different setup and conditions were met in the second stage, test activities and data collection follows. This procedure provides details as to how the data was collected in the room setup, the simulation activity for the source of pollutants, and the timeframe needed for the device to collect data. When all the necessary data under different conditions were collected, the raw data undergoes the stage of preprocessing and data splitting. Raw data was secured to be clean, and normalized, making the data classification-ready, and undergoes data splitting for the training proper. After the previous procedure, data training follows using a pattern recognition technique identified in this study. The trained data was then validated based on its accuracy and performance evaluation was done through the given statistical methods involved in this research.

#### A. Research Conceptualization

The ground theory of the previous literature builds the foundation of this research. The literature of this study revolves around (1) the significance of why IAQ is needed and the possible health impact on humans, (2) the indoor air quality research trends, giving details to literature who were well-studied, researched, and studies with limited resource contribution, (3) IAQ parameters to consider using the sensor results and readings under different thermodynamic conditions,

and (4) different pattern recognition techniques based on the machine learning research trends.

In the last several years, indoor air quality has been a well-studied area in the environmental research community. Different environmental organizations have taken steps to establish standards and guidelines to address issues and potential risks to human health. Technological advancements are also given attention to the research community and IAQ industry on intelligent systems, IoT, machine learning, etc. During the literature review, limited research contribution was seen in the area of indoor air quality which involves pattern recognition techniques. This identified gap motivated the researcher to conduct further study on the topic. The study of Saad Shaharil Mad et al. [19], was the main literature that gave motivation to this research. This study is an improvement to the study of Saad Shaharil Mad et al. [19] by taking into consideration the use of targeted parameters under different thermodynamic conditions.

1) *Research approach and initial strategies:* The study's general objective was to classify five (5) indoor air pollutants using thermal, particle, and gas parameters under two (2) thermodynamic conditions using multilayer perceptron pattern recognition technique. Thus, the research approach should be quantitative. Another consideration in selecting what research approach to use was based on the typical approach used by the related literature in this research. Also, the nature of the research objectives should clearly define what strategy to utilize. Since the research is about the classification process in machine learning then the strategy to pursue should be the predictive strategy.

2) *Participants and setting:* The usual strategy for the collection of data regarding air quality uses environmental chambers and other controlled room setups. In the study of Shisheng, Chen et al. [23] regarding time series prediction of CO<sub>2</sub>, TVOC, and HCHO, they made use of the rooms with the air-conditioning unit (ACU) inside the campus of the National University of Singapore (NUS) as the location for the air quality data collection [23]. In the same way, this research used the facility of the University of the Philippines – Cebu as the location for collecting air parameters in classifying pollutant sources. Specifically, this research was conducted in the Department of Computer Science Conference Room, 3rd floor, Room 313 of the Arts and Sciences (AS) building University of the Philippines - Cebu.

#### B. Experimental Setup and Specifications

Before acquiring the data for preprocessing and data training, needed preparations were done to achieve organized procedures regarding the simulated test activities. This subpart talks about the setup and conditions inside the room, pollutant sources and parameters to consider, and the IAQ device monitor to use.

1) *Room setup and conditions:* Fig. 2 below shows the room's details, fixtures, and specifications. The figure was the floorplan venue for the collection of indoor air parameters. The room has the dimension of 5.33m x 3.10m x 3.05m, it has one door, three windows, one window-type air-conditioning unit (ACU), and fixtures like couches (small and big), a conference table with office chairs, and small cabinet.

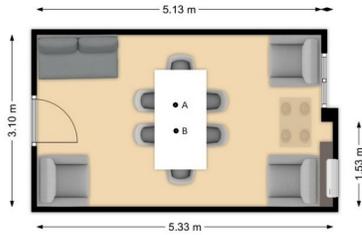


Fig. 2. Room specifications using floorplanner.com

2) *Sources of indoor air pollutants:* The research also adopted one of the indoor air pollutants activities being used in the study of Demanega, Ingrid, et al. [11] which was candle burning for combustion activity. Other source pollutants adopted came for the study of Saad, Shaharil Mad et al. [19] such as cleaning agents like “Lysol”, rotten cooked fish, and the ambient condition. This research also included rubbing alcohol as one of the sources which is currently a big contributor to indoor air pollutants because of the pandemic. Table I shows the summary of the source pollutants included in this research with the activity description, data collection interval, and span. Each pollutant undergoes two thermodynamic conditions cool & dry and warm & humid. Data was then collected every 1-minute interval and a collection span of 16 hours or 960 samples will be collected across each of the source pollutants considering the thermodynamic condition. An 8-hour day plan was decided in consideration of the university’s working hours. Gathering samples for the 5 single source pollutants for each thermodynamic condition gives us a total of 9600 samples or 160 hours of collection time. Given an 8-hour day plan, 160 hours is equal to 20 days collection process.

TABLE I. ACTIVITY DESCRIPTION, DATA COLLECTION INTERVAL OF THE SOURCE POLLUTANTS

Source Pollutants	Thermodynamic Condition	Activity Description	Data Collection Interval
Ambient	Cool and Dry	Observe the ambient condition reading of the surrogate sensors and device reference.	Collect data every 1 minute. A total of 960 samples should be collected.
	Warm and Humid		
Combustion	Cool and Dry	Light up two candles in the room.	
	Warm and Humid		
Chemical	Cool and Dry	Place half full glass of isopropyl alcohol in the middle of the room.	
	Warm and Humid		
Fragrance	Cool and Dry	Automatic Spray using Lysol for every 18 minutes in the room.	
	Warm and Humid		
Rotten Food	Cool and Dry	5 days old rotten food(leftovers) will be placed in the middle of the room	
	Warm and Humid		

3) *Indoor air quality parameter and device monitor:* In the study of Saad Shaharil Mad et al. [19], nine parameters were used to develop a pollutant recognition based on pattern recognition techniques, but the study of Demanega Ingrid et al. [11] paved way to consider only using a few target parameters in classifying indoor pollutants. That same

study gave results showing positive detection outcomes for all sensors and devices. Most of the sensors responded to the simulated indoor pollutant sources which gave way to adopting the previous studies’ parameters. The pollutant parameters that were considered were air temperature and relative humidity for thermal parameters, particulate matter (PM2.5/10) for particle parameters, and carbon dioxide as a gas parameter. Thus, this study considers the parameters that were studied in the previous related work.

The study of Demanega Ingrid et al. [11], provides an assessment of low-cost monitors, research, and professional-grade IAQ systems used as reference monitors. These systems are expensive, yet the reading accuracy is top-notch with respect to the different IAQ parameters. Not just expensive, most of these systems detect a single parameter only unlike the consumer-grade IAQ monitors which are available in the market. But in that same study, some of the consumer-grade monitors were assessed and exhibited good performance grading in detecting the different parametric values for pollutant sources. On that note, the IAQ monitor that was used in this study was low-cost and consumer-grade monitors targeting the relevant parameters that were considered in the previous study. Two IAQ handheld devices were considered in implementing this research. Temptop M2000 2nd Generation was used to collect CO2 and particulate matter. BR-SMART-123SE was used to collect TVOC and HCHO. These devices were then positioned in the middle of the room, specifically on the top of the table.

4) *Working with thermodynamic conditions:* According to the study of Xiangguo et al. [24], conventional all-air central air-conditioning (AC) systems can control the temperature and humidity through cooling, reheating, and humidifying equipment but AC systems which were commonly seen in small and medium-sized buildings have no specific dehumidifying equipment to deal with moisture. In the Philippines, relative humidity is high because of the surrounding body of water. It is said on the PAGASA website that the average monthly relative humidity varies between 71 percent in March and 85 percent in September. In a requirement to have fresh air indoors, good ventilation by opening the windows can help but will greatly influence the moisture level. In working with thermodynamic conditions and achieving the cool & dry and warm & humid room setup, doors and windows were sealed to avoid moisture air influencing the indoor humidity and temperature. Also, dehumidifiers and humidifier equipment were placed inside to control the humidity level of the room while an AC unit was used to control the room temperature with a significant contribution to the relative humidity. The dehumidifier and humidifier equipment has automatic control which directly reacts with the varying humidity levels in the room.

### C. Data Collection and Test Activities

Before the collection procedure, the two thermodynamic conditions should be first met. In the case of cool and dry, the ACU was turned on together with the dehumidifier for some time until the conditions were met for an activity to happen. For relative humidity control, only the dehumidifier was turned on since the room space was not completely sealed. A humidity level of 30% was never achieved in consideration with the study of Demanega Ingrid et al. [11]. Only 45% +/- 5% was

achieved in this research. Both the ACU and dehumidifier have a control device that shuts off when a certain temperature or humidity value is reached. For warm and humid, ACU will be turned off and both the dehumidifier and humidifier were turn on to control the humidity level of the room. In the same way, the ACU, dehumidifier, and humidifier were connected to a control device that establishes the right thermodynamic condition. When the conditions were set, different source pollutants were injected. Also, the device monitors which were placed in the middle of the room were turned on at least 1 hour before each activity allowing the sensors to have enough time to stabilize and the collection procedure commences as soon as the desired thermodynamic conditions were met.

Table I shows the order in collecting the source pollutants data with the desired activity description. A warm and humid condition will be implemented first to be followed by a cool and dry condition. The start and end of the collection procedure will be manually timestamped to map the desired data activity. PM2.5, CO2, TVOC and HCHO was collected while considering thermodynamic conditions. Thus, giving this research a total of 5 parameters.

#### D. Data Preprocessing, Exploration and Splitting

The raw dataset that was then generated from the device monitor and was extracted through USB in excel format. The manual timestamp was defined and divided into different data based on the thermodynamic conditions and different pollutant sources. After organizing the dataset, data cleaning was followed. This was where incorrect data type or format, missing values, and data duplication in the dataset were either modified, replaced, or deleted before the data normalizatin/standardization process and choosing other data preprocessing techniques. Preprocessing is important because it may aid in the success of pattern recognition performance [19].

After the procedures above, data exploratory analysis follows. This process will investigate the data values and explore meaningful insights. One process for data exploration is through descriptive statistics which gives this research the first insight for interpretation or an overview of what the clean raw data look like. This process will also give insight into the potential outlier readings inside the dataset which can be deleted. After doing descriptive statistics, this research will explore some of the questions identified below.

1. What are the relationships between the collected parameters?
2. Are the collected values different considering two thermodynamic conditions?

Answering these questions through the process of data exploratory analysis gives the researcher an initial understanding of what the dataset looks like. The initial analysis taken through the dataset was then generated using visual representation. Data exploratory analysis use python software to generate the visual results.

After the data exploratory analysis, data splitting was followed by dividing the cleaned and explored raw data into a training dataset and a testing dataset. The training dataset was used for the classification training and the testing dataset was used to check how the current generated model was working.

If the model was not performing well, an iterative process of training to develop a pattern recognition model followed by testing dataset checking occurs. The testing dataset will be used to evaluate the classification model's performance. 60% of the data will be taken for training and 40% will be taken for the testing procedure.

#### E. Data Training and Testing

According to the study of Saad Shaharil Mad et al. [19], an artificial neural network (ANN) was more suitable to be embedded in an IAQ system. The algorithm does not need large storage space unlike the other counterparts and is easy to embed because it requires a less complicated formula. ANN are parallel information processing approaches that are applied for data processing, process analysis, control, fault detection, pattern recognition, defining the complex and non-linear relationship, and employing a number of input-output training patterns from the experimental data [25]. The commonly used neural network architecture is the multilayer feed-forward neural network known as Multilayer Perceptrons or MLP networks which are based on a backpropagation algorithm and comprise multiple hidden layers and neurons. Adding one or more hidden layers creates another set of synaptic connections and more neural interactions which leads to the improvement of the network's accuracy. Fig. 3 shows the general architecture of the multilayer feed-forward neural network for prediction.

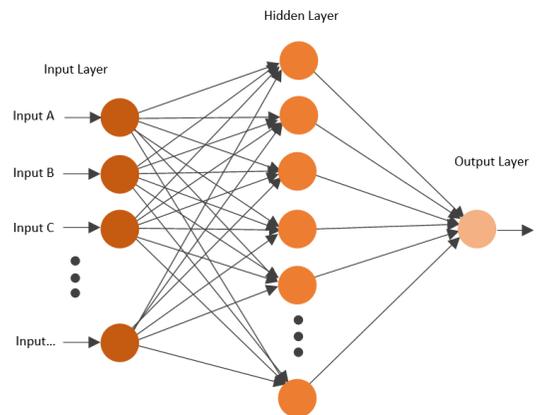


Fig. 3. Sample multilayer perceptron network architecture

Hyperparameters in the study of Saad Shaharil Mad et al. [19] were adopted in this study. One hidden layer was used with three (3) neurons. Vector array normalization was used as feature scaling and an assumption of using stochastic gradient descent was used as the optimizer algorithm since there was no mention in the previous study. Other hyperparameters like the input layer, batch size, kernel initializer, activation function in the last layer, etc. are all defined based on the nature of the classification problem. Using the hyperparameters, training of the dataset will follow to generate the multilayer perceptron model accuracy and loss function. Optimization was then carried out to improve the model with the consideration of overfitting and underfitting. GridSearchCV was then used to find the values of the optimal hyperparameters.

F. Performance Evaluation and Software Details

Evaluation of the performance of the classification model was done through model training and validation metric of accuracy while the loss function was through categorical cross-entropy.

The high-level programming language that was used to implement the methods regarding data cleaning, exploration, normalization, splitting, training, testing, and evaluation was done through Python language. This research utilized the integrated development environment (IDE), Spyder, as the open-source platform for scientific programming in the python language. During the implementation using the platform, software libraries imported such as pandas, sci-kit-learn, Keras, and TensorFlow, etc.

IV. RESULTS AND DISCUSSION

In this study, the IAQ handheld devices were set up in the middle of the room to collect PM2.5, CO2, HCHO, and TVOC. Secondly, temperature and humidity equipment were prepared to achieve two thermodynamic conditions commonly encountered in many climates around the world. During the preparation of the experimental room setup, a specific limitation was found before the implementation of the data collection process. Before injecting a source pollutant, the two thermodynamic conditions must be met first. In the case of the cool and dry condition, room temperature should be at 20C +/- 1C and humidity of 30% +/- 5%. It was found that the room condition can reach the required temperature for the cool and dry condition but can never reach the humidity value of 30% +/- 5%. This limitation was brought about by the experimental room setup which was not completely sealed. This setup simulates the regular room setup where pollutant sources can be found. The final humidity value of 45% +/- 5% was set for cool and dry condition. Finally, adding the limitation above, the simulation of different indoor air pollutants was administered while collecting the data values generated by the devices. This section reports the results and findings in relation to the research question and hypothesis of this study while providing meaning, importance, and relevance of the result.

A. Preliminary Analysis

Initially, data cleaning was performed to the collected raw dataset to ensure correctness and improve data quality. The cleaned raw dataset was processed through preliminary analysis by describing key features of the data. Each parameter for all categorical conditions was correlated to provide insights into the parameter's relationship. Table II shows the individual categorical correlation.

Examining the table above shows that for ambient conditions, it was observable that HCHO has a very high positive correlation with respect to the thermodynamic condition and has a value of 0.92. For combustion conditions, PM2.5 and TVOC has negative high relationship with respect to thermodynamic conditions. For chemical conditions, HCHO has a very high positive correlation with the thermodynamic condition

TABLE II. CATEGORICAL PEARSON CORRELATION

Class 0	Ambient Condition				Class 1	Combustion Condition					
	PM2.5	CO2	HCHO	TVOC		PM2.5	CO2	HCHO	TVOC		
PM2.5	1.000				PM2.5	1.000					
CO2	0.400	1.000			CO2	0.310	1.000				
HCHO	0.087	0.560	1.000		CO2	-0.170	0.055	1.000			
TVOC	0.220	0.480	0.340	1.000	CO2	0.450	0.200	0.300	1.000		
CD/WH	-0.023	0.480	0.920	0.049	1.000	CO2	-0.790	-0.077	0.220	-0.710	1.000
Class 2	Chemical Condition				Class 3	Fragrance Condition					
	PM2.5	CO2	HCHO	TVOC		PM2.5	CO2	HCHO	TVOC	CD/WH	
PM2.5	1.000				PM2.5	1.000					
CO2	0.062	1.000			CO2	0.600	1.000				
HCHO	0.370	0.079	1.000		CO2	0.590	0.690	1.000			
TVOC	0.390	0.320	0.670	1.000	CO2	0.200	0.429	0.680	1.000		
CD/WH	0.390	0.084	0.930	0.590	1.000	CO2	0.430	0.670	0.800	0.830	1.000
Class 4	Rotten Food Condition										
	PM2.5	CO2	HCHO	TVOC	CD/WH						
PM2.5	1.000										
CO2	0.410	1.000									
HCHO	-0.130	0.020	1.000								
TVOC	0.110	0.190	0.960	1.000							
CD/WH	-0.200	-0.059	0.960	0.910	1.000						

and a moderate correlation value of 0.59 with respect to TVOC. For the fragrance conditions, most of the parameters has a moderate to very high positive correlation with respect to each other. Lastly for rotten food conditions, HCHO and TVOC has a high positive correlation with thermodynamic conditions.

The categorical dataset was then merged to provide insight to have the overall Pearson correlation of the dataset's parameters including the categorical conditions. Fig. 4 shows the heatmap generated through seaborn in python.



Fig. 4. Overall correlation heatmap

Firstly, the figure above shows that PM2.5 has a high positive correlation value with respect to CO2 which measures a positive correlation of 0.77 which verifies the result in most of the IAP conditions having a significant correlation value. The same positive correlational value was also observed for HCHO and TVOC parameters. Strong correlation was found for HCHO and TVOC since HCHO was part of the parameters to be collected in TVOC device. Secondly, CO2 has negligible correlational value with respect to the thermodynamic condition and has a very minimal negative correlation with respect to TVOC and HCHO with the value of 0.37 and 0.2 respectively. The result of CO2 is also verified using the correlation table of the individual IAP condition. Thirdly, the thermodynamic condition has a low correlational value to HCHO of positive 0.32 which verifies as well in the individual IAP condition correlation table. Lastly, the overall correlation of the IAP conditions has a very low negative correlation with respect to PM2.5 and CO2 with negligible correlation to other variables.

**B. Accuracy and Loss Comparison between Distinct Scaling Technique and Optimizer Algorithm**

After the preliminary analysis, model training was initiated. The study of Saad Shaharil Mad et al. [19] model training algorithm motivates this study in adopting multilayer perceptron as the only machine learning algorithm and adopted some of the hyperparameters used in the previous study. Table III shows the adopted hyperparameters and using additional hyperparameters to carry out this study.

TABLE III. MULTILAYER PERCEPTRON HYPERPARAMETERS

Adopted Hyperparameters	
Hidden Layer	1
Neurons	3
Output Layer	5
Learning Rate	0.01
Momentum Constant	0.5
Epochs	1000
Additional Hyperparameters	
New Input Layer	5
Batch Size	64 (based on a paper)
Feature Scaling	VAN, Standardization
Optimizer Algorithm	SGD, Adam
Hidden Layer Activation	Relu (default on Keras)
Kernel Initializer	glorot_uniform (default on Keras)
Output Layer Activation	Softmax
Metric	Accuracy
Loss Function	Categorical Cross Entropy

This study addresses the hypothesis by using only 5 IAQ parameters in predicting the different classes of indoor air pollution, thus, setting the input layer equal to 5. The hidden layer, neurons, output layer, learning rate, and momentum constant were directly adopted in generating the model accuracy and loss. Additional hyperparameters were set to successfully train the data using Multilayer Perceptron. This study uses a fixed batch size of 64, in reference to the recommendation of the study of Kandel, Ibrahim et al [26]. For feature scaling, Vector Array Normalization was seen to be the best performer in the previous research, and standardization was not included for comparison. This study has drawn insights into the difference in results between using VAN and Standardization (STD) technique. Optimizer Algorithm was never mentioned in the previous research; thus, this study assumes and adopted the use of stochastic gradient descent (SGD) as the optimizer for this research. Also, this study has explored the Adam optimizer since this optimizer is always compared with SGD in many papers. For this comparison setup, this study has used 'relu' for the hidden layer activation, 'glorot\_uniform' for the kernel initializer, and 'softmax' for the output layer activation function. The two former variables were considered as default hyperparameter of Keras for model training. 'Softmax' was used as the output layer activation function since this is an example of a multiclass classification problem.

After setting up the hyperparameters, model training starts. Accuracy and loss function results was compared using the combination of two feature scaling and two optimizer algorithms.

Given Fig. 5, in using VAN as the feature scaling and SGD as the optimizer, it can be observe that the accuracy output

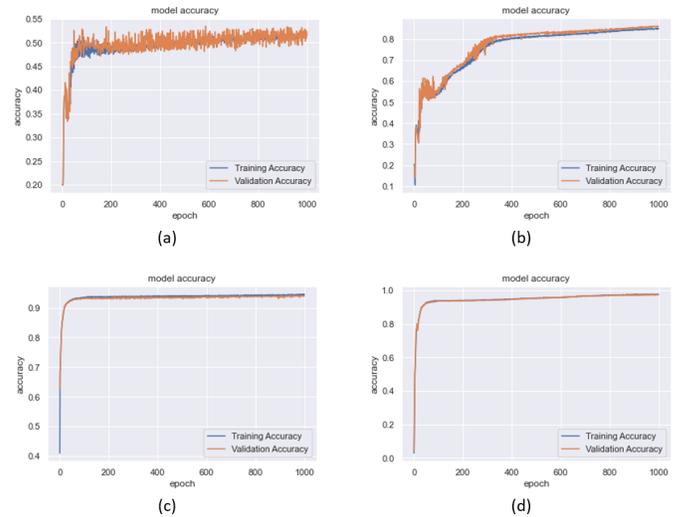


Fig. 5. Model training and validation accuracy (a) VAN-SGD (b) VAN-Adam (c) STD-SGD (d) STD-Adam

was noisy, and the training process was slow reaching to a validation accuracy of 52% for 1000 epochs. Using the same feature scaling while changing SGD to Adam optimizer, the accuracy results are less noisy, and the validation accuracy improves reaching 86% for 1000 epoch. Using standardization (STD) as the feature scaling while varying the optimizer, the result greatly improves. With SGD as the optimizer, the validation accuracy incurs a value of 93.8% compared to Adam optimizer with a value of 97% for 1000 epochs.

Fig. 6 shows the model loss comparison. Model loss in VAN feature scaling was observed to be bigger compared to the feature scaling of standardization after 1000 epochs. VAN-Adam has better loss value compared to VAN-SGD with 0.34 and 0.99 validation loss respectively. Both optimizers using STD as the feature scaling reaches a loss value of less than 0.2 in 1000 epochs. STD-SGD incur a validation loss of 0.16 compared to STD-Adam with a validation of 0.07.

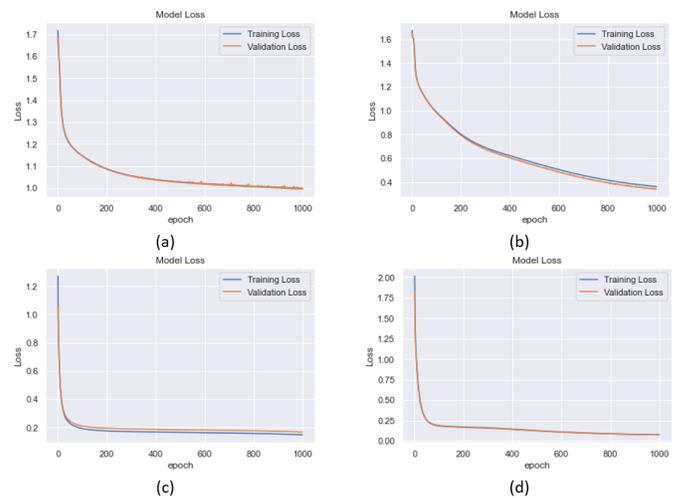


Fig. 6. Model training and validation loss (a) VAN-SGD (b) VAN-Adam (c) STD-SGD (d) STD-Adam

### C. Hyperparameter Tuning

Referring to the comparison results in the above section, the accuracy in using STD and Adam can be seen as good fit already. Even if the result was good, it can be observed that some of the hyperparameters was not tuned with the hypothesis of gaining the optimized accuracy and loss value. To further improve the results, tuning of different hyperparameters was investigated. Hyperparameters like number of input layers, number of hidden layers, number of neurons, number of output layers, output layer activation function, metrics and loss function were preset because of the nature of the dataset. The result in the previous section in using standardization as the feature scaling and Adam as the optimizer algorithm was also included as a preset value for the tuning process. Momentum was excluded in this process because of the new optimizer. It was also observed that in using STD as feature scaling, 1000 epochs was too much of a training period and eventually just limits the number of epochs to 50. Table IV shows the validation accuracy and loss function for the different epochs. The table shows that the use of standardization (STD) feature scaling is superior compared to vector array normalization (VAN) in this dataset. It is observable that even if the epoch was 50, using standardization still exhibits a good accuracy and loss value.

TABLE IV. VALIDATION RESULT PER EPOCHS

	Validation Result					
	1000 epochs		100 epochs		50 epochs	
	Accuracy	Loss	Accuracy	Loss	Accuracy	Loss
VAN-SGD	0.5216	0.9952	0.4794	1.1598	0.4711	1.2210
VAN-Adam	0.8617	0.3420	0.5706	1.0109	0.4599	1.1576
STD-SGD	0.9383	0.1169	0.9284	0.2206	0.9102	0.2902
STD-Adam	0.9711	0.0750	0.9362	0.1837	0.9255	0.2475

Taking into consideration the execution time of model training, hyperparameters like learning rate and batch size was varied. Other hyperparameters like hidden layer activation and kernel initializer was also varied to find the best possible combination of hyperparameters. Table V shows the summary of hyperparameters.

TABLE V. NEW SET OF HYPERPARAMETERS

Hyperparameters	
Input Layer	5
Hidden Layer	1
Neurons	3
Output Layer	5
Learning Rate	0.1, 0.01, 0.001
Epochs	50
Batch Size	32, 64, 128
Feature Scaling	Standardization (STD)
Optimizer Algorithm	Adam
Hidden Layer Activation	Tanh, Relu, Sigmoid
Kernel Initializer	glorot_uniform, uniform, he_uniform
Output Layer Activation	Softmax
Metric	Accuracy
Loss Function	Categorical Cross Entropy

Given all the hyperparameters above, model training was initiated and using python's gridsearchCV to exhaustively

search for the best fit combination of hyperparameters. Cross validation (CV) for this process was set to 5.

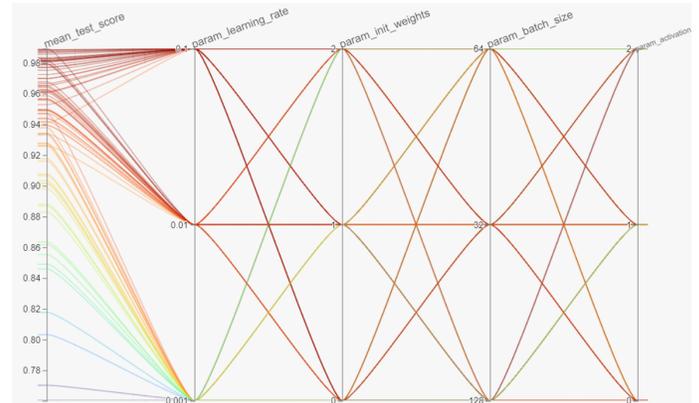


Fig. 7. GridSearchCV hyperparameter combinations

Fig. 7 shows the different hyperparameter combinations with the mean test score value where mean test score is the mean accuracy of the classification. Using STD and Adam optimizer as the new hyperparameter, it is shown that all hyperparameter combination exhibits the value of mean test score greater than 70%. It also shows that there are hyperparameter combinations that reaches more than 99% of mean test score.

Table VI shows top 10 performing hyperparameter combinations with respect to the mean test score with mean fit time

TABLE VI. TOP PERFORMING HYPERPARAMETER COMBINATION BASED-ON MEAN TEST SCORE

Activation	Batch Size	Initial Weights	Learning Rate	Mean Test Score	Mean Fit Time
Relu	32	be_uniform	0.01	0.98924	15.98910
Sigmoid	64	glorot_uniform	0.1	0.98854	8.60077
Relu	64	Uniform	0.1	0.98819	8.35799
Sigmoid	32	glorot_uniform	0.1	0.98901	15.08160
Sigmoid	32	be_uniform	0.1	0.98698	14.59695
Sigmoid	32	Uniform	0.1	0.98681	15.14623
Tanh	128	glorot_uniform	0.1	0.98559	4.98741
Relu	128	glorot_uniform	0.1	0.98333	4.68067
Sigmoid	128	glorot_uniform	0.1	0.98264	4.75186
Relu	128	be_uniform	0.1	0.98264	4.82595

It was observable that rank 1 combination garnered a mean test score of 98.9 percent while having a mean fit time of 15.878 seconds. Succeeding combinations have good mean test score values but has longer mean fit time. The mean fit time is the average time of training between cross validation folds. The correlation figure below shows the relationship between mean test score and mean fit time with the varying hyperparameters.

TABLE VII. CORRELATION OF VARYING HYPERPARAMETERS WITH MEAN TEST SCORE AND MEAN FIT TIME

	Activation	Batch Size	Initial Weights	Learning Rate
Mean Test Score	-0.1963	-0.2720	0.0035	0.5749
Mean Fit Time	0.0212	-0.9222	0.0194	-0.0091

Table VII shows the correlation of varying hyperparameter results. Batch size can be seen to have a high negative correlation with respect to the mean fit time which gives an idea that training period is greatly influence by the value of batch size. The bigger the batch size, the lesser the mean fit time and the lower the batch size, the bigger the mean fit time. Negligible relationship can be seen through learning rate, activation function and initial weights with respect to the mean fit time. Moderate positive correlation value was found from mean test score with respect to learning rate and very low to negligible relationship can be seen from mean test score with respect to hidden layer activation and batch size and initial weight.

#### D. Accuracy and Loss using the Optimal Hyperparameter based on Mean Test Score

Using the optimal value of hyperparameters based on mean test score, model training was again initiated to get the accuracy and loss function. Fig. 8 shows the model training and validation results.

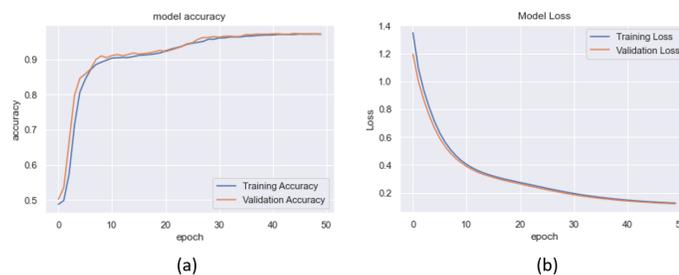


Fig. 8. Model training and validation using the optimal hyperparameter(a) accuracy (b) loss

Taking into consideration Table IV and Fig. 8, it is observable that the accuracy and loss function in 50 epochs greatly improves using the tuned hyperparameters in comparison to old hyperparameters. Accuracy of the old hyperparameters in 50 epochs was 92.55% while the new hyperparameter incurred a value 98.9%. Loss function for old hyperparameter in 50 epochs was 0.2475 in comparison of the new hyperparameter with a value of 0.0986.

#### V. CONCLUSION

This research presents the process of classifying indoor air pollutant sources using relevant parameters under two thermodynamic conditions. The research question that was formulated at the start of this study was finally addressed using this research. Model accuracy and loss function values were seen to be in good fit which is comparable to the previous study. From nine parameters, this research streamlines the number of parameters into five considering CO<sub>2</sub>, PM<sub>2.5</sub>, TVOC, HCHO, and Thermodynamic Conditions. Using the dataset generated in this study feature scaling was compared using normalization and standardization. Standardization scaling was shown to be superior in relation to the vector array normalization scaling which was used in the previous study. The hyperparameters

used in the previous study were also taken into consideration for this research to have the optimal values of hyperparameters. It can be observed that hyperparameters should be tuned to gain optimal result value. Computational time was greatly reduced at the same time giving the best-fit accuracy and loss function values without incurring the problem of overfitting.

The current study classifies indoor pollutants with a minimal number of sensors but can only classify one pollutant source at a time. The previous study collected a dataset on a single source and tested the acquired model on mixed pollutant sources but relatively gains poor performance. Considering this limitation, further dataset gathering and investigation on mixed pollutant sources should be examined while taking into consideration hyperparameter tuning.

#### ACKNOWLEDGMENT

The researcher would like to thank Central Visayas Studies Center (CVSC) of University of the Phillipines Cebu for funding this research. Also, special thanks to Lovelie and Elle for being an inspiration.

#### REFERENCES

- [1] P. O. Fanger, *What is IAQ?*, DOI: 10.1111/j.1600-0668.2006.00437.x, 2006
- [2] R. Pitarma, G. Marques and B. R. Ferreira, *Monitoring Indoor Air Quality for Enhanced Occupational Health*, J Med Syst 41, 23 (2017). <https://doi.org/10.1007/s10916-016-0667-2>
- [3] M. J. Jafari, A. A. Khajevandi, S. IM. Najarkola, M. S. Yekaninejad, M. A. Pourhoseingholi, L. Omidi, S. Kalantary, *Association of Sick Building Syndrome with Indoor Air Parameters*, 2015; 14(1):55-62. PMID: 26221153; PMCID: PMC4515331.
- [4] United States Environmental Protection Agency (USEPA), *emphIndoor Air Quality*, <https://www.epa.gov/indoor-air-quality-iaq/introduction-indoor-air-quality>
- [5] J. Saini, M. Dutta, G. Marques, *A comprehensive review on indoor air quality monitoring systems for enhanced public health*, Sustain Environ Res 30, 6 (2020). <https://doi.org/10.1186/s42834-020-0047-y>
- [6] V. V. Tran, D. Park, Y. C. Lee, *Indoor Air Pollution, Related Human Diseases, and Recent Trends in the Control and Improvement of Indoor Air Quality*, Int J Environ Res Public Health. 2020;17(8):2927. Published 2020 Apr 23. doi:10.3390/ijerph17082927
- [7] World Health Organization, Regional Office for Europe, *Air quality guidelines for Europe, 2nd ed.. World Health Organization*, <https://apps.who.int/iris/handle/10665/107335>, 2000
- [8] M. Mannan, S. G. Al-Ghamdi, *Indoor Air Quality in Buildings: A Comprehensive Review on the factors Influencing Air Pollution in Residential and Commercial Structure*, Int. J. Environ. Res. Public Health 2021, 18(6), 3276; <https://doi.org/10.3390/ijerph18063276>
- [9] M. Tastan, H. Gokosan, *Real-time Monitoring of Indoor Air Quality with Internet of Things- Based E-nose*, Appl. Sci. 2019, 9(16), 3435; <https://doi.org/10.3390/app9163435>
- [10] G. Chiesa, S. Cesari, N. Garcia, M. Issa, S. Li, *Multisensor IoT Platform for Optimising IAQ levels in Buildings Through a Smart Ventilation System*, Sustainability 2019, 11(20), 5777; <https://doi.org/10.3390/su11205777>
- [11] I. Demanega, I. Mujan, B. C. Singer, A. S. Andelkovic, F. Babich, D. Licina, *Performance assessment of low-cost environmental monitors and single sensor under variable indoor air quality and thermal conditions*, 2021, Building and Environment 187 (2021) 107415
- [12] Z. Wang, W. W. Delp, B. C. Singer, *Performance of low cost Indoor air quality monitors for PM2.5 and PM10 from residential sources*, <https://doi.org/10.1016/j.buildenv.2020.106654>
- [13] W. Yang, J. Sohn, J. Kim, B. Son, J. Park, *Indoor air quality investigation according to age of the school buildings in Korea* Journal of Environmental Management 90 (2009) 348e354

- [14] C. Mandin, M. Trantallidi, A. Cattaneo, N. Canha, V. G. Mihucz, T. Szigeti, R. Mabilia, E. Perreca, A. Spinazzè, S. Fos-sati, Y. Kluzenaar, E. Cornelissen, I. Sakellaris, D. Saraga, O. Hänninen, E. O. Fernandes, G. Ventura, P. Wolkoff, P. Car-ner, J. Bartzis, *Assessment of indoor air quality in office build-ings across Europe – The OFFICAIR study*, Science of The Total Environment, Volume 579, 2017, Pages 169-178, ISSN 0048-9697, <https://doi.org/10.1016/j.scitotenv.2016.10.238>.
- [15] A. Datta, R. Suresh, A. Gupta, D. Singh, P. Kulshrestha, *Indoor air quality of non-residential urban buildings in Delhi, India*, International Journal of Sustainable Built Environment, Volume 6, Issue 2, 2017, Pages 412-420, ISSN 2212-6090, <https://doi.org/10.1016/j.ijse.2017.07.005>.
- [16] D. Saraga, T. Maggo, E. Sadoun, E. Fthenou, H. Hassan, V. Tsiouri, S. Karavoltos, A. Sakellari, C. Vasilakos, K. Kakosimos, *Chemical Characterization of Indoor and Outdoor Particulate Matter (PM<sub>2.5</sub>, PM<sub>10</sub>) in Doha Qatar*, <https://doi.org/10.4209/aaqr.2016.05.0198>
- [17] M. Castelli, F. M. Clemente, A. Popovič, S. Silva, L. Van-neschi, *A Machine Learning Approach to Predict Air Quality in Cal-ifornia*, Complexity, vol. 2020, Article ID 8049504, 23 pages, 2020. <https://doi.org/10.1155/2020/8049504>
- [18] W. Wei, O. Ramalho, L. Malingre, S. Sivanantham, J. C. Little, C. Mandin *Machine learning and statistical models for predicting indoor air quality*. Indoor Air. 2019 Sep;29(5):704-726. doi: 10.1111/ina.12580. Epub 2019 Jul 25. PMID: 31220370.
- [19] S. M. Saad, A. M. Andrew, M. Shakaff, M. Dzahir, M. Hussein, M. Mohamad, Z. A. Ahmad, *Pollutant Recognition Based on Supervised Machine Learning for Indoor Air Quality Monitoring Systems* Appl. Sci. 2017, 7, 823. <https://doi.org/10.3390/app7080823>
- [20] J. Kallio, J. Tervonen, P. Räsänen, R. Mäkynen, J. Koivusaari, J. Peltola, *Forecasting office indoor CO<sub>2</sub> concentration using machine learning with a one-year dataset*, Building and Environment, Volume 187, 2021, 107409, ISSN 0360-1323, <https://doi.org/10.1016/j.buildenv.2020.107409>.
- [21] H. Zhang, R. Srinivasan, V. Ganesan, *Low-Cost , Multi-Pollutant Sensing System Using Raspberry Pi for Indoor Air Quality Monitoring*, Sustainability 2021, 13, 370. <https://doi.org/10.3390/su13010370>
- [22] American Society of Heating Refrigerating and Air-Conditioning En-gineers - ASHRAE, *Thermal environmental conditions for human occu-pancy*, ANSI/ASHRAE Stand. - 55. 7 (2017) 6.
- [23] S. Chen, K. Mihara, J. Wen, *Time series prediction of CO<sub>2</sub>, TVOC and HCHO based on machine learning at different sampling points*, Building and Environment, Volume 146, 2018, Pages 238-246, ISSN 0360-1323, <https://doi.org/10.1016/j.buildenv.2018.09.054>.
- [24] X. Xiangguo, S. Chao, D. Shiming, *Using two parallel connected compressors to implement a novel control algorithm for improved indoor humidity at a low cost* Building Services Engineering Research and Technology. 2013;34(3):349-354. doi:10.1177/0143624412446092
- [25] A. Rostami, M. Anbaz, H. Gahrooei, M. Arabloo, A. Bahadori, *Accurate estimation of CO<sub>2</sub> adsorption on activated carbon with multi-layer feed-forward neural network (MLFNN) algorithm*, Egyptian Journal of Petroleum, Volume 27, Issue 1, 2018, Pages 65-73, ISSN 1110-0621, <https://doi.org/10.1016/j.ejpe.2017.01.003>.
- [26] I. Kandel, M. Castelli, *The effect of batch size on the generalizability of the convolutional neural networks on a histopathology dataset*, ICT Express, Volume 6, Issue 4, 2020, Pages 312-315, ISSN 2405-9595, <https://doi.org/10.1016/j.ict.2020.04.010>.